# DEVELOPMENT OF REGIONAL SKEW MODELS FOR RAINFALL FLOODS IN CALIFORNIA USING BASEYIAN LEAST SQUARES REGRESSION

A Thesis

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Master of Science

by

Jonathan Richard Lamontagne

January 2014

# ABSTRACT

The thesis here reports on and expands the results published in Lamontagne et al. [2012]. A hybrid Bayesian weighted/generalized least squares regression procedure is used to generate regional skew models for annual maximum rainfall floods of various durations in California. The procedure uses weighted least squares to estimate the model coefficients, and generalized least squares to estimate model precision. This procedure is necessitated by the unusually high cross-correlation exhibited between concurrent rainfall floods at different sites, which caused the regression weights to become unjustifiably erratic. New diagnostic statistics are developed for this special case and applied to real data. Overall model precision is excellent, which is important in the context of Bulletin 17B flood frequency analysis.

Chapter 1 of the thesis provides an introductory background to flood frequency analysis, and the scope and area of the study. Chapter 1 also describes the procedure used by the United States Army Corps of Engineers to develop the rainfall flood time series.

Chapter 2 discusses the characteristics of the log-Pearson Type III distribution, the Bulletin 17B flood frequency procedure, the Expected Moments Algorithm, and the effect of outliers on frequency estimation and tests for their identification and removal.

Chapter 3 describes the development of weighted least squares and generalized least squares for regionalization of hydrologic variables. Chapter 3 then derives the new hybrid weighted/generalized least squares regression procedure and its

accompanying diagnostic statistics. Finally, Chapter 3 discusses recent research which uses an alternative generalized least squares framework.

Chapter 4 details the application of the procedure from Chapter 3 to rainfall flood of various durations from California to create a regional skew model for California.

Finally, Chapter 5 examines various aspects of the analysis in Chapter 4 which were noticeably different from previous regional skew studies. In particular, Chapter 4 reexamines the Pseudo ANOVA table and proposes a new, alternative table.

## BIOGRAPHICAL SKETCH

Jonathan Richard Lamontagne was born on November 11, 1986 in Nashua, NH to Marc George Lamontagne and Pamela Anne Lamontagne. In 2009 he received a B.S. in Civil Engineering from the University of New Hampshire. In the fall of 2009 he entered Cornell and began work with Dr. Jery Stedinger towards an MS/PhD in Civil and Environmental Engineering. Jonathan's current research areas include flood frequency analysis and hydropower operations optimization. During the summer of 2011, Jonathan married Katelyn, Louise Trexler

For Mom, Dad, and my amazing wife Katie

ACKNOWLEDGMENTS

I would like to sincerely thank Dr. Jery Stedinger for his countless hours of attention and his enduring patience. His enthusiasm for research and dedication to quality are unsurpassed. I would also like to thank Dr. Wilfred Brutsaert and Dr. Huseyin Topaloglu for their guidance in selecting courses and for serving on my Committee.

I am deeply grateful to the United States Geological Survey and the United States Army Corps. Of Engineers for their initial support of this research and valuable insights during its initial stages.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

# PREFACE

This Thesis focuses on flood frequency analysis, and more specifically on the application of Bayesian Generalized Least Squares regression for the generation of regional skew models. The specific application presented here is the generation of regional skew coefficient models for California Rainfall Floods of five durations. The genesis of this work was a collaborative regional skew study between the US Geological Survey, the US Army Corps of Engineers, and researchers at Cornell University. That work is available as a US Geological Survey Scientific Investigation Report (Lamontagne et al., 2012), available through the California Water Science Center. This thesis provides a more in depth and extended discussion of the analysis presented there.

Chapter 1 contains general background information about the motivation and scope of the study. Section 1.1 details the background, including a brief discussion of flood frequency procedures in the United States and current improvements to California infrastructure that precipitated this work. Section 5.2 discusses the geographic scope of the study area, which mostly included river basins which drain into the Central Valley of California. Section 1.3 details the range of basin characteristics which were available for each study basin as well as a discussion as to their range across the study region. Section 1.4 describes the flood separation procedure used by the US Army Corps of Engineers when developing the rainfall flood records, and Section 1.5 discusses several notable basins which required special treatment.

Chapter 2 contains a discussion of flood frequency analysis procedures based on the log-Pearson type III distribution. Section 2.1 explores the characteristics of the

log-Pearson type III distribution, and Section 2.2 details the Bulletin 17B fitting procedure, which is the standard procedure applied by Federal agencies in the United States. Section 2.3 describes the Expected Moments Algorithm, which is a new moments based fitting technique which efficiently accounts for outliers and special data types. Finally, Section 2.4 describes the problem of low outliers in flood frequency analysis, and explores several procedures for their identification.

Chapter 3 provides a theoretical discussion of Bayesian GLS procedures generally, and a description of the new Bayesian WLS/GLS procedure applied by Lamontagne et al. [2012]. Section 3.3 also introduces the concept of redundant basins and provides a new statistic designed to detect potentially redundant pairs.

Chapter 4 discusses the details of the regional skew analysis for California rainfall floods first published by Lamontagne et al. [2012]. Section 4.6 provides an auxiliary analysis, not published in Lamontagne et al. [2012], which explores whether a unique model form for each duration would have been appropriate.

Chapter 5 explores several issues which were raised during the US Geological Survey internal review process. Section 5.1 explores the variance of prediction and the effective record length statistics to determine if the remarkable results in Chapter 4 are appropriate when compared to other Bayesian GLS skew studies. Section 5.2 explores the issue of Analysis of Variance and discusses a new addition to the pseudo Analysis of Variance proposed by previous Bayesian GLS work. Finally, Section 5.3 explores whether the unique regional skew models for each duration result in inconsistencies in the subsequent flood frequency analysis.

# CHAPTER 1

# INTRODUCTION

When designing civil infrastructure, engineers must consider the natural loadings which a structure will experience over its lifetime. For riparian structures, this often involves estimating the magnitude of a flood associated with a certain frequency, or return period, or the risk of flooding. Flood frequency analysis attempts to estimate the frequency of flood magnitudes based on flood records, basin hydrologic characteristics, and a combination of both. Many methods can be utilized in flood frequency analysis. This thesis focuses on the *Bulletin 17B* (B17B) guidelines, which are the standard flood frequency methodology used by US Federal agencies. More specifically, this thesis focuses on estimation of the regional log-space skew coefficients, which are combined with at-site sample skew coefficients to fit the log-Pearson Type III (LP3) distribution to annual maximum flood series [Interagency Committee on Water Data (IACWD), 1982; Stedinger et al., 1993; Griffis and Stedinger, 2007a; England and Cohn, 2008]

The research presented here builds on previous work in the Bayesian Generalized Least Squares (GLS) regional skew coefficient regression framework [Reis et al., 2005; Gruber and Stedinger, 2008]. Methodological advancements include the first application of a newly developed least squares algorithm designed to accommodate high cross-correlation among the sample skew coefficients, and the introduction of a new redundancy metric, standardized distance. The focus of this

study is rainfall floods of various durations in and around the Central Valley of California.

Section 1 of this chapter contains a discussion of flood frequency methods that are commonly applied in the United States, recent developments, and the purpose and scope of this research. Section 2 describes the hydrology of the study region and the characteristics of the basins which were selected for inclusion. Section 3 describes the basin characteristics considered as explanatory variables for regional skew, and Section 4 briefly describes the procedure used to estimate annual rainfall floods. Finally, Section 5 describes special considerations taken for notable study basins.

### *Section 1.1: Background*

The risk of flooding is a major consideration for riparian structures. Obtaining reliable estimates of flood frequency is a very important design consideration. If one can assume that a gauged basin's hydrology has been stationary over some period of record, then the frequency of a flood magnitude can be estimated directly from the flood record. Unfortunately, the length of such records is often limited compared to the long return periods of interest, so direct determination of the underlying frequency distribution is not precise enough to be of practical use. For example, one would not expect a 30-year record to directly yield a good estimate of the 100-year flood. In some applications, such as dam design, engineers estimate 1,000-year flood events [Calzascia and Fitzpatrick, 1989], while the maximum record length for a basin in the United States is about 110 years. Fitting a statistical distribution to flood records provides a way to extend the frequency information contained in a record to extreme events which have likely not been observed. For this reason, rigorous flood frequency

analysis often involves the fitting of statistical distributions to historical flood records [IACWD, 1982; Stedinger et al., 1993; Brutsaert, 2005; Chow et al., 1988].

Several distributions are reasonable for describing annual flood series, including the log-normal (LN), log-Pearson Type III (LP3), and the generalized extreme value (GEV) distributions [Stedinger et al., 1993, Chow et al., 1988]. One can imagine that extreme flood quantiles estimated from different frequency distributions can yield very different flood flow estimates, which can in turn have result in very different flood stage estimates.

Prior to the 1960s, no uniform method for flood frequency analysis had been adopted by the United States Federal government, which is responsible for the construction and maintenance of most critical infrastructure in the United States. Up to that time individual agencies each used their own preferred methods, which often led to inconsistencies. In an effort to encourage more consistency, the recently formed Water Resources Council published *Bulletin 15* "A Uniform Technique for Determining Flood Flow Frequencies" in 1967. *Bulletin 15* dictated that the LP3 distribution with a regional skew coefficient be used by all Federal agencies when conducting flood frequency analysis. This was followed by *Bulletin 17* in 1976, *Bulletin 17A* in 1977, and finally *Bulletin 17B* (B17B) in 1981, with slight revisions released in 1982. These later releases provided further guidance on fitting the LP3 to increase uniformity across all agencies [Griffis and Stedinger, 2007b; Griffis, 2006; IACWD, 1982]. The LP3 distribution and specific B17B recommendations are discussed in more detail in Chapter 2.

Since 1982 there have been no official changes to B17B, but research on flood frequency using the LP3 has continued. Recent advances can be divided into five primary categories[Stedinger and Griffis, 2008]: regional skew estimation, use of historical information, plotting positions, confidence intervals, and quantile weighting. This thesis is focused on improving techniques for regional skew estimation, and represents the latest development in regional skew regression methodology. Another recent advance has been improved outlier detection methodologies, which will be discussed tangentially in Chapter 2.

B17B provides a national skew map (Plate I) for estimation of regional skew, but also recommended the analyst seek better, regional models when possible [IAWCD, 1982]. To this end, Tasker and Stedinger [1986] proposed a Weighted Least Squares (WLS) procedure for regressing regional skew coefficient models on basin characteristics. This framework assumes regression error to be caused by either model error or sampling error of the sample skew coefficient. Sites are weighted according to the sampling error variance of their sample skew coefficient (a function of record length). Stedinger and Tasker [1985] and Tasker and Stedinger [1989] laid out a Generalized Least Squares (GLS) procedure for hydrologic regression. This differs from a WLS framework in that GLS also accounts for the covariance between observations. The significance of this is discussed in more detail in Chapter 3. Reis et al. [2005] extended this procedure to a Bayesian GLS framework for regional skew models. Traditional GLS can return zero model error variance in some instances, which is clearly not realistic. By contrast, Bayesian GLS generates the posterior distribution of model error variance, so this is not a concern. Bayesian GLS has been

applied to regional skew coefficient modeling in the southeast region of the United States [Feaster et al., 2009; and Gotvald et al., 2009; Weaver et al., 2009] and in California [Parrett et al., 2011], with additional studies currently on going.

Kjeldsen and Jones [2006, 2007, 2009] have also applied GLS to generate index flood models for the United Kingdom. Kjeldsen and Jones [2007] present evidence that model errors are correlated, while Stedinger and Tasker [1985] assume they are not. The hydrologic regionalization work of Kjeldsen and Jones are described in more detail in Section 3.3and by Veilleux [2009].

The purpose of this study was to develop regional skew coefficient models for rainfall floods of various durations in California using the advances described above. This was accomplished using a new WLS/GLS hybrid method similar to that used in the previous California instantaneous annual peak skew study [Parrett et al., 2011]. This study was the result of collaboration between the United States Army Corps of Engineers (USACE), the California Department of Water Resources (DWR), the United States Geological Survey (USGS) California Water Science Center (CAWSC), and the School of Civil and Environmental Engineering at Cornell University.

As is the case around the United States, much of the flood protection infrastructure in California is in dire need of repair and is slated for rehabilitation or reconstruction [ASCE, 2006; Lamontagne et al., 2012]. Before embarking on this effort in earnest, the USACE sought to develop the most rigorous study possible of flood risk in the Central Valley of California, using the most advanced techniques available. USGS Water Science Centers provide periodic updates to flood frequency estimates at sites of interest, so CAWSC and USACE collaborated in this effort.

Given the success of the instantaneous annual peak flow study [Parrett et al., 2011], USACE sought to expand the study to include annual maximum rainfall floods of 1-day, 3-day, 7-day, 15-day, and 30-day durations. Since researchers at Cornell University had developed the most recent advances in regional skew regression and successfully collaborated with USGS to conduct regional skew studies in the Southeast [Feaster et al., 2009; and Gotvald et al., 2009; Weaver et al., 2009] and California [Parrett et al., 2011], USGS approached Cornell to conduct the regional skew coefficient aspect of the study [Lamontagne et al., 2012].

The USACE provided Cornell with annual maximum unregulated rainfall flood records for each of the durations of interest at 55 basins in and around the Central Valley of California. In some cases these had to be reconstructed from regulation records using the process discussed in Section 1.4. The USGS provided key basin characteristics for consideration as explanatory variables. These are discussed in more detail in Section 1.3.

### *Section 1.2: Study Area*

The site records selected for this study are primarily from dams and gauging stations on rivers and streams which flow into California's Central Valley and are operated by USACE or USGS respectively. USACE provided records for 55 sites, but only 50 were used in this study. Site records were removed for a variety of reasons including flood record unreliability, flood records that were synthetically augmented using other study sites, or because a site experienced flows which were uncharacteristic of the greater study region. An example of this last point was Orestimba Creek near Newman (study site 22), which experienced 12 years of zero

flow in a record of 77 years, with several other reported maximum annual flows being less than 5 cfs. The flood distribution characteristics of Orestimba Creek were deemed too different from the greater study region and not characteristic of the critical infrastructure sites of interest to USACE.

Table 1.1 lists the site number and name, the available period of flood record (POR), the mean basin elevation (Elev), and drainage area (DA) for each site included in the study. Figure 1.1 shows the location of each of the study sites and their basins overlaid on a map of California.

The study sites' basins cover a wide range of hydrologic types and can be divided into three geographical categories: Sierra Nevada Range basins, north Coastal Range basins, and south Coastal Range basins. Both high mountain basins which receive deep annual snow pack and small flat basins on the Central Valley floor were included. Mean basin elevations included in this study range between 250 and 7,500 ft, with drainage areas between 10 and 6,400 sq miles. Figure 1.2 and Figure 1.3 plot the drainage area and mean basin elevation versus site respectively.

**Table 1.1:** Study basins, period of record, drainage area, and mean basin elevation [Reproduced from Lamontagne et al., 2012].

| Site # | Site Name | POR | DA | Elev (ft) |
|---|---|---|---|---|
| 1 | Sacramento R Shasta Dam | 77 | 6403 | 4571 |
| 3 | Cottonwood Ck near Cottonwood | 68 | 922 | 2221 |
| 4 | Cow Ck near Millville | 59 | 423 | 2251 |
| 5 | Battle Ck below Coleman Hatch | 68 | 361 | 4074 |
| 6 | Mill Ck near Los Molinos | 80 | 131 | 3962 |
| 7 | Elder Ck near Paskenta | 60 | 93 | 2998 |
| 8 | Thomes Ck at Paskenta | 76 | 204 | 4146 |
| 9 | Deer Ck near Vina | 92 | 209 | 4199 |
| 10 | Big Chico Ck near Chico | 77 | 72 | 3111 |
| 11 | Stony Ck at Black Butte Dam | 66 | 740 | 2416 |
| 12 | Butte Ck near Chico | 78 | 148 | 3717 |
| 13 | Feather R At Oroville Dam | 107 | 3591 | 5031 |
| 14 | North Yuba at Bullards Dam | 68 | 489 | 4899 |
| 15 | Bear R near Wheatland | 103 | 292 | 2250 |
| 16 | N Fork Cache Ck at IV Dam | 77 | 120 | 2627 |
| 17 | American R at Fair Oaks | 104 | 1887 | 4356 |
| 18 | Kings R at Pine Flat Dam | 113 | 1544 | 7634 |
| 19 | San Joaquin R at Friant Dam | 105 | 1639 | 7046 |
| 20 | Chowchilla R at Buchanan Dam | 80 | 235 | 2152 |
| 23 | Del Puerto Ck near Patterson | 44 | 73 | 1835 |
| 24 | Merced R at Exchequer Dam | 107 | 1038 | 5473 |
| 25 | Tuolumne R at Don Pedro Dam | 112 | 1533 | 5882 |
| 26 | Stanislaus R at Melones Dam | 93 | 904 | 5663 |
| 28 | Duck Ck near Farmington | 30 | 11 | 249 |
| 30 | Calaveras R at Hogan Dam | 96 | 372 | 1991 |
| 31 | Mokelumne R at Camanche Dam | 104 | 628 | 4918 |
| 32 | Cosumnes R at Michigan Bar | 101 | 535 | 3064 |
| 33 | Fresno R near Knowles | 76 | 134 | 3201 |
| 34 | S Yuba R at Jones Bar | 57 | 311 | 5362 |
| 35 | M Yuba R below Our House Dam | 37 | 145 | 5365 |
| 36 | Kaweah R at Terminus Dam | 50 | 560 | 5635 |
| 37 | Tule R at Success Dam | 50 | 392 | 3975 |
| 38 | Kern R Isabella Dam | 116 | 2075 | 7198 |
| 39 | Mill Ck near Piedra | 52 | 115 | 2637 |
| 40 | Dry Ck near Lemoncove | 50 | 76 | 2668 |
| 41 | Deer Ck near Fount Spring | 41 | 83 | 3989 |
| 42 | White R near Ducor | 46 | 91 | 2443 |
| 43 | Cache Ck at Clear Lake | 87 | 527 | 2004 |
| 44 | Putah Ck at Mont Dam | 78 | 567 | 1327 |
| 45 | M Fork Eel R near Dos Rios | 43 | 745 | 3685 |
| 46 | S Fork Eel R near Miranda | 68 | 537 | 1726 |
| 47 | Mad R above Ruth Res | 28 | 94 | 3705 |
| 48 | E Fork Russian R near Calpella | 67 | 92 | 1630 |
| 49 | Salinas R near Pozo | 41 | 70 | 2211 |
| 50 | Arroyo Seco near Soledad | 107 | 241 | 2494 |
| 51 | Salmon R at SomesBar | 84 | 751 | 4261 |
| 52 | Santa Cruz Ck near Santa Ynez | 67 | 74 | 3355 |
| 53 | Salsipuedes Ck near Lompoc | 67 | 47 | 920 |
| 54 | Trinity R above Coffee Ck | 51 | 148 | 5340 |
| 55 | Scott R near Fort Jones | 67 | 662 | 4333 |

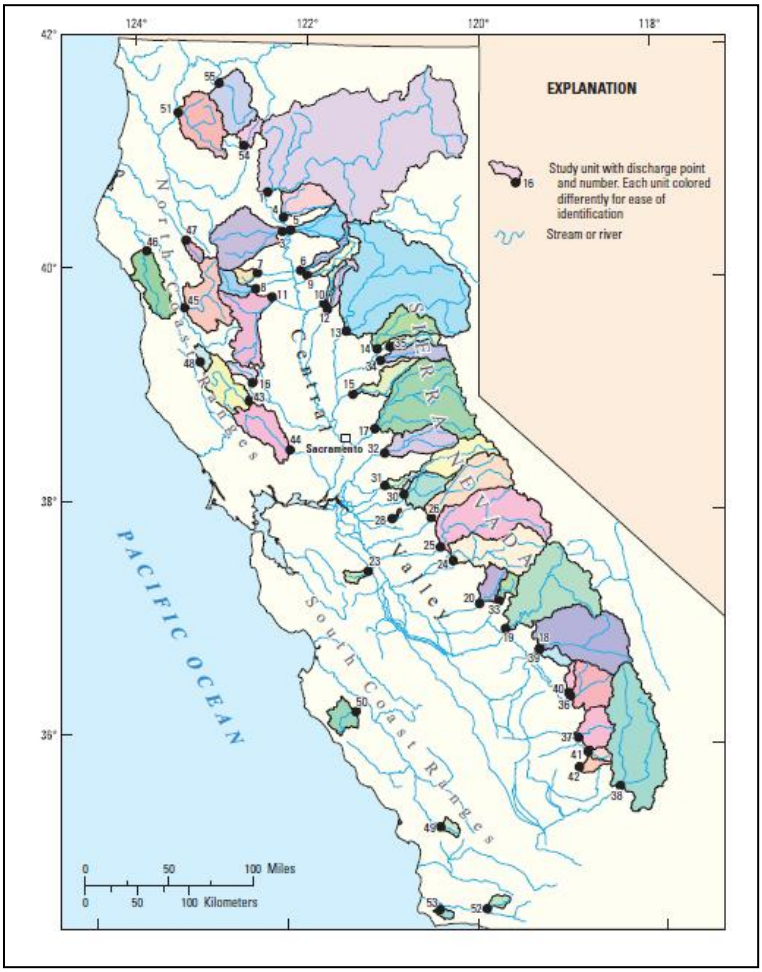


**Figure 1.1:** Location of Study Basins [Reproduced from Lamontagne et al., 2012].

The majority of basins in the study drain the western side of the Sierra Nevada Mountain Range, which runs along California's eastern border. These basins generally range in mean elevation between 2,000 and 7,500 ft, and have drainage areas roughly between 90 and 6,000 sq miles. The southern Sierra Nevada basins, including the Kern River (study site 38), the Kaweah River (study site 36), the Kings River (study site 18), and the Tule River (study site 37) drain some of the highest elevations in the continental US, including Mt. Whitney, which is the highest peak in the lower 48 states [Carle, 2004].

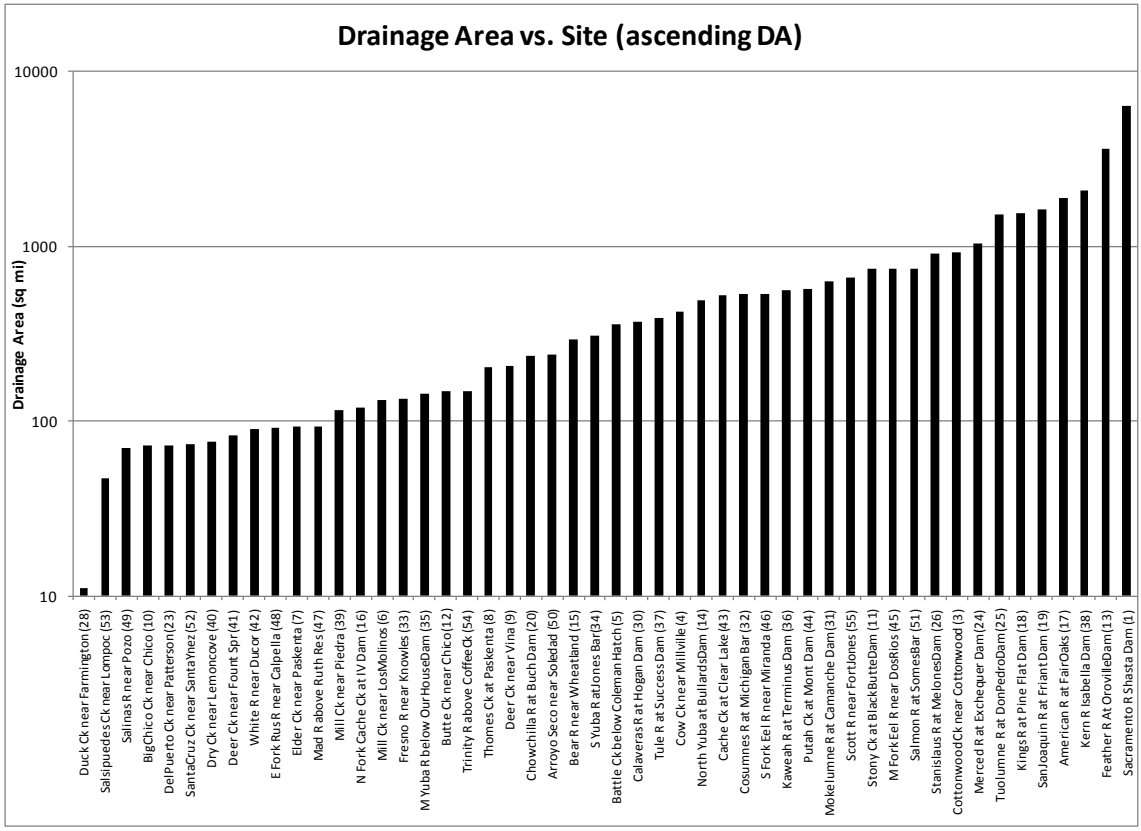


**Figure 1.2:** Drainage Area versus Site [Reproduced from Lamontagne et al., 2012].

The basins of the central Sierra Nevada flow into the San Joaquin River, which account for roughly 9% of California's total annual runoff. Study basins in this region include eastern sites between the Cosumnes River at Michigan Bar (study site 32) in the north to the upper San Joaquin River at Friant Dam (study site 19) in the south (see Figure 1.1). The Cosumnes River is the only Sierra Nevada river in this study which is not regulated. The major tributaries to the lower San Joaquin River are the Stanislaus River (study site 26), Toulmne River (study site 25) and the Merced River (study site 24) [Carle, 2004]. These basins represent some of the highest mean elevation basins in the study, matched only by the southern Sierra Nevada study basins.

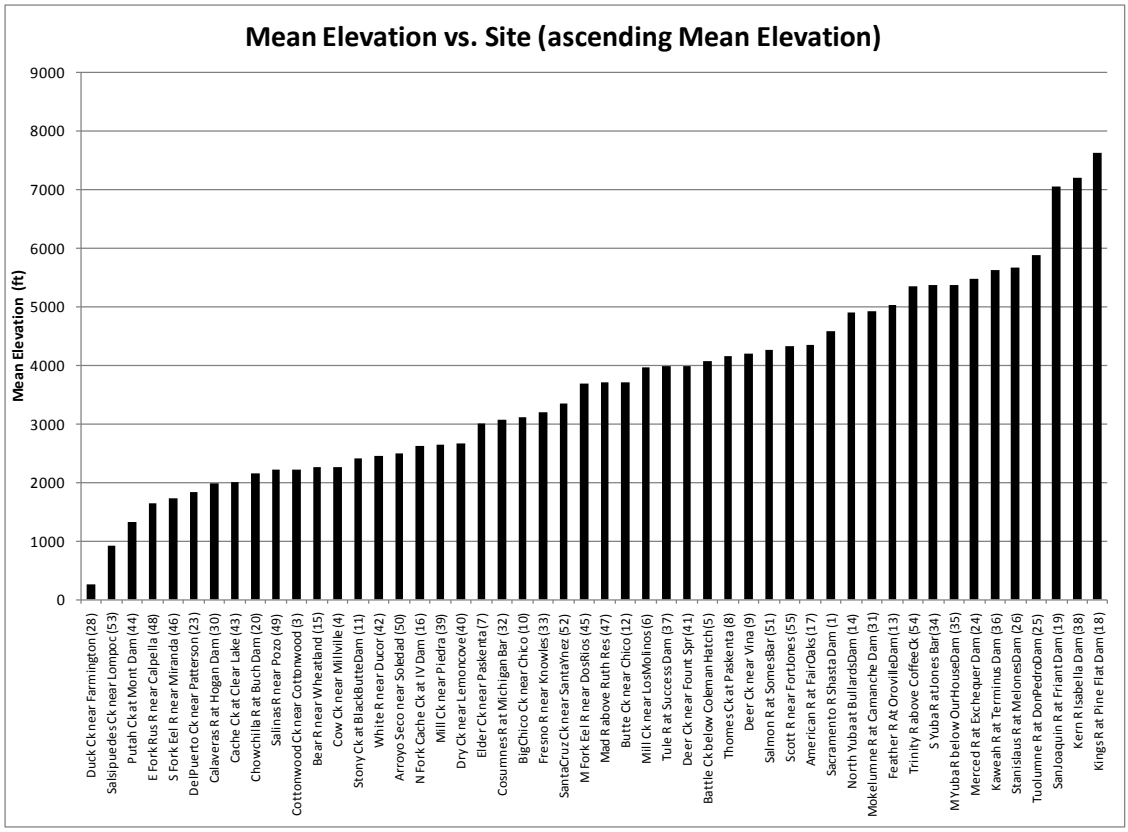


**Figure 1.3** Mean Basin Elevation vs. Site [Reproduced from Lamontagne et al., 2012].

The basins of the northern Sierra Nevada region contribute the majority of the flow of the Sacramento River, which drains over 30% of California's total annual runoff [Carle, 2004]. In this study the Mt. Shasta region, including the Sacramento River at Shasta Dam (study site 1), the Trinity River above Coffee Creek (study site 54), and the Scott River at Fort Jones (study site 55) were considered Sierra Nevada basins.

Many of the Sierra Nevada study basins experience significant snowfall in the winter months and prolonged snowmelt runoff in the spring. This study dealt with rainfall floods only, so exclusive snowmelt events were not considered. However, the largest annual floods in these basins are often rain-on-snow events in which warming temperatures cause precipitation to fall as rain, which rapidly melts much of the

snowpack, resulting in large flood events. This phenomenon was observed by Parrett et al. [2011], who conducted a more extensive instantaneous annual peak flow skew study for California. Special considerations taken for these events are described in Section 1.4.

The majority of the other basins included in the study drain both the eastern and western sides of the Coastal Range, which runs roughly parallel to the Pacific coast from Oregon to the Mexican border. In this study, the Coastal Range was divided into two groups, with basins north of San Francisco forming the north Coastal Range and basins south of San Francisco forming the south Coastal Range.

The north Coastal Range experiences the highest rainfalls in California, with many basins experiencing over 100 inches of rain per year. This region accounts for roughly 40% of the annual runoff in California [Carle, 2004]. Major study basins in this region include the Eel River (both Middle (45) and South (46) Forks), the Salmon River gauged at Somes Bar (study site 51), and the East Fork of the Russian River gauged near Capella (study site 48). Incredibly, the Eel River has produced flood flows which exceed those experienced on the Sacramento River despite its drainage area being less than one-seventh the size. This is an indication that extreme rainfall events dominate the north Coastal Range flood characteristics. Some higher elevation north Coastal Range basins do experience some snowfall in the winter months, but this is relatively limited compared to the Sierra Nevada snow packs [Carle, 2004].

It should be noted that three sites draining the interior of the north Coastal Range were included in this classification: Cache Creek at Clear Lake (study site 43), North Fork Cache Creek at the Indian Valley Dam (study site 16), and Putah Creek at

Monticello Dam (study site 44). These sites experienced many small rainfall floods during their period of record, and are not necessarily characteristic of the other north Coastal Range sites described above. Their unique hydrology required special consideration when performing flood frequency as discussed in Section 2.4.

The south Coastal Range basins experience much less rainfall than the north Coastal Range basins. Major basins in this region include Santa Cruz Creek near Santa Ynez (study site 52) and the Salinas River near Pozo (study site 59). Santa Cruz Creek originates in mountainous terrain and has the highest mean basin elevation of all the south Coastal Range basins. The Salinas River often dries out due to evaporation, infiltration, and diversion, but feeds groundwater aquifers that sustain local agriculture. This is not uncommon for this region. The aforementioned Orestimba Creek, whose record contained 12 zero annual rainfall flood observations from a record of 77 years, is located in this region. The Arroyo Seco River (study basin 50) is the only major river in the south Coastal Region which is not regulated [Carle, 2004].

Figure 1.4 plots the mean basin elevation versus drainage area for the three geographical categories of sites described above. Note that the largest and highest basins in the study are located in the Sierra Nevada Range, though many Sierra Nevada basins have similar mean elevation and drainage area as Coastal basins. North Coastal basins tend to have larger drainage areas than south Coastal basins, though both have similar mean basin elevations. The sites with the largest drainage areas in this study are the Sacramento River at Shasta Dam (study site 1) followed by the Feather River at Oroville Dam (study site 13). The three study sites with the highest mean basin elevation are the Kings River at Pine Flat Dam (study site 18), the Kern

River at Isabella Dam (study site 38), and the San Joaquin River at Friant Dam (study site 19). These basins are located in the south of the Sierra Nevada Range. In both the Coastal and the Sierra Nevada Mountain Ranges, southern basins tend to receive less annual precipitation than northern basins [Carle, 2004], and often more snow because of the higher elevation.

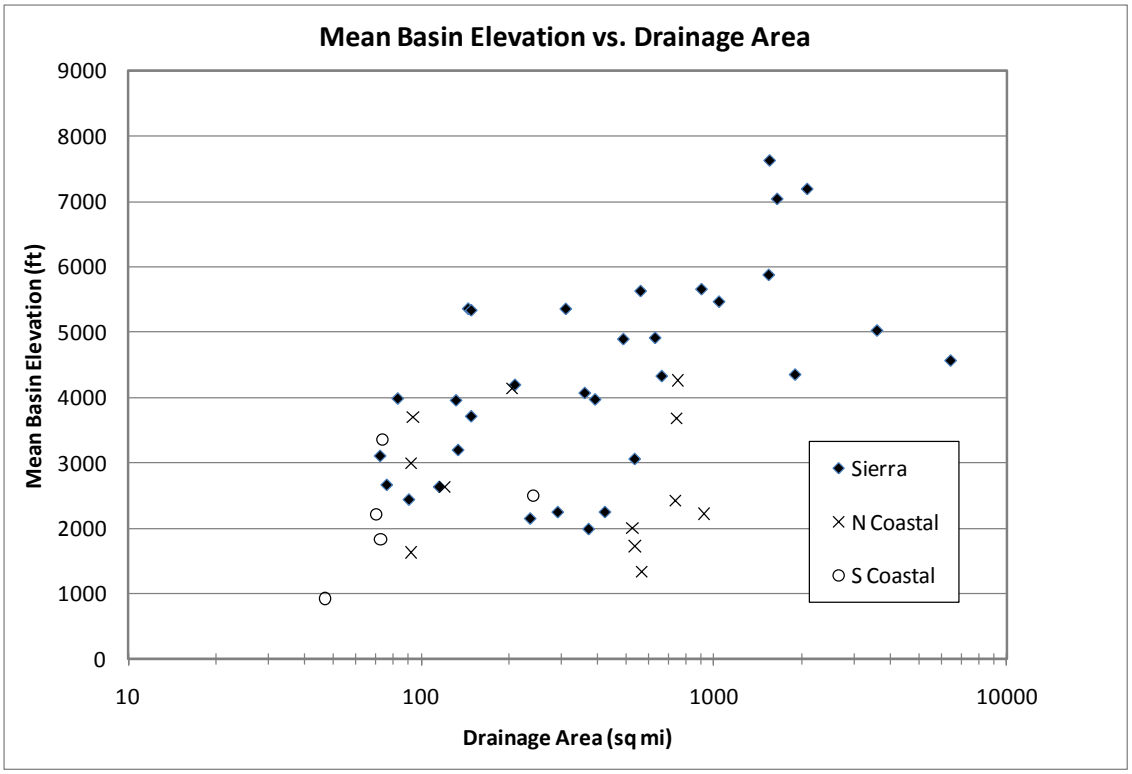


**Figure 1.4:** Drainage Area vs. Mean Basin Elevation (Study Site 28 not Plotted) [Reproduced from Lamontagne et al., 2012].

The basin with the smallest drainage area and mean basin elevation is Duck Creek near Farmington (study site 28), which is located on the floor of the Central Valley. It is the only study site not plotted in Figure 1.4. This basin has a drainage area of 11 square miles and a mean basin elevation of roughly 250 ft. While Duck Creek is not necessarily characteristic of the basins of interest to USACE, its short

record length (only 30 years) ensured that it had very little weight in the skew regression.

### *Section 1.3: Basin Characteristics*

A total of 20 basin characteristics were provided by USGS for each of the study sites. These included common physical characteristics such as mean basin elevation and basin drainage area, geographic characteristics such as basin centroid and outlet location, and climatic characteristics such as mean annual precipitation and mean January maximum temperature. Geographic and physical characteristics were drawn from the National Hydrologic Dataset (NHDPlus) and National Land-Cover Dataset (NLCD). Climatic characteristics were largely drawn from the Parameter-Elevation Regressions on Independent Slopes Model (PRISM) climatic dataset. Climatic data was compared to the older National Water Information System (NWIS) database and inconsistencies were found to be negligible [Lamontagne et al., 2012]. Table 1.2 contains a list of the basin characteristics considered in this study, along with definitions, and the source of the data.

Of particular significance to this study was mean basin elevation, EL6000, and basin centroid location. Mean basin elevation is simply the mean elevation above sea level in feet and the centroid location is specified by its latitude and longitude in decimal degrees. EL6000 was the percent of the basin area with elevation greater than 6000 feet, which is generally considered the elevation above which winter precipitation falls as snow in California. Each of these basin characteristics proved to be significant through the course of the study, as discussed in Chapter 4.

**Table 1.2:** List of all primary variables considered [Reproduced from Lamontagne et al., 2012].

| Name | Description | Data source (if applicable) |
|---|---|---|
| Basin No | Unique identifier for basin | |
| BASINPERIM | Perimeter, in miles | 30-m DEM, NHDPlus elev_cm grid http://www.horizon-systems.com/NHDPlus/ |
| RELIEF | Relief, in feet | 30-m DEM, NHDPlus elev_cm grid http://www.horizon-systems.com/NHDPlus/ |
| Mean ELEV | Average basin elevation, in feet | 30-m DEM, NHDPlus elev_cm grid http://www.horizon-systems.com/NHDPlus/ |
| DRNAREA | Basin drainage area, in square miles | |
| ELEVMAX | Maximum elevation, in feet | 30-m DEM, NHDPlus elev_cm grid http://www.horizon-systems.com/NHDPlus/ |
| MINBELEV | Minimum elevation, in feet | 30-m DEM, NHDPlus elev_cm grid http://www.horizon-systems.com/NHDPlus/ |
| LAKEAREA | Percentage of area covered by lakes and ponds | 2001 National Land Cover Database (NLCD) - Land Cover http://www.mrlc.gov/nlcd_multizone_map.php |
| EL6000 | High Elevation Index - Percent of basin area with elevation above 6,000 feet | 30-m DEM, NHDPlus elev_cm grid http://www.horizon-systems.com/NHDPlus/ |
| OUTLETELEV | Elevation at outlet, in feet | 30-m DEM, NHDPlus elev_cm grid http://www.horizon-systems.com/NHDPlus/ |
| RELRELF | Basin relief divided by basin perimeter, in feet per mile | |
| DIST2COAST | Distance in miles from basin centroid to coast along a line perpendicular to eastern California border | |
| BSLDEM30M | Average basin slope, in percent | 30-m DEM, NHDPlus elev_cm grid http://www.horizon-systems.com/NHDPlus/ |
| FOREST | Percentage of basin covered by forest | 2001 National Land Cover Database (NLCD) - Percent Canopy http://www.mrlc.gov/nlcd_multizone_map.php |
| IMPNLCD01 | Percentage of basin covered by impervious surface | 2001 National Land Cover Database (NLCD) - Percent Impervious http://www.mrlc.gov/nlcd_multizone_map.php |
| PRECIP | Mean annual precipitation, in inches | 800M resolution PRISM 1971-2000 data http://www.prism.oregonstate.edu/products/ |
| JANMAXTMP | Average maximum January temperature, in degrees Fahrenheit | 800M resolution PRISM 1971-2000 data http://www.prism.oregonstate.edu/products/ |
| JANMINTMP | Average minimum January temperature, in degrees Fahrenheit | 800M resolution PRISM 1971-2000 data http://www.prism.oregonstate.edu/products/ |
| CENTROIDX | X coordinate of the centroid, in decimal degrees | |
| CENTROIDY | Y coordinate of the centroid, in decimal degrees | |
| OUTLETX | X coordinate of the basin outlet | |
| OUTLETY | Y coordinate of the basin outlet | |

### *Section 1.4: Procedures for Determining Rainfall-floods*

Time series of unregulated annual maximum rainfall floods of the five durations were prepared by USACE for each of the 50 study sites. This section provides a brief explanation of the process by which these were obtained. A more in

depth discussion of this can be found in USACE [2002] and Lamontagne et al. [2012].

In this study, daily maximum flows are measured from midnight-to-midnight rather than in a moving 24-hour window. Unregulated daily flow data were available for 28 of the 50 study sites, while 22 study sites experience some regulation or diversions. The objective in this study was to develop regional skew coefficient models for annual maximum unregulated rainfall flood flows. Thus, USACE had to determine and remove the effects of regulation and snowmelt. In general, a four step methodology was used to achieve this:

(1) Obtain a daily streamflow record for basin.

(2) If necessary, consider the daily regulation record and remove effects of regulation. In some cases this involved observing the daily change in storage record for a basin's terminal reservoir.

(3) Observe daily hydrograph of large events and remove snowmelt-only events.

(4) Extract remaining annual maximum flow.

The process for event separation involves visual inspection of the daily flow series supplemented with daily temperature data to determine the start of the snowmelt season. This then serves as the segregation point for events. If the maximum event for the year occurs due to rainfall during the snowmelt period, the segregation point was adjusted to include this event in the rainfall floods.

Twenty-two of the study sites' records exhibited significant snowmelt effects so that the maximum recorded flood in many years was caused snowmelt only. These events were identified through visual inspection of the flood record and removed

[Lamontagne et al. 2012]. Each of these sites are located in the Sierra Nevada Range, which experiences much more snowfall than the Costal Range.

Twenty-seven sites in this study were included in the previous Sacramento-San Joaquin Comprehensive Study, which featured annual maximum rainfall flood values through 1998 or 1999 [USACE, 2002]. These records were extended through 2008 using the procedure discussed above. The Calaveras River at New Hogan Dam (study site 30) was also extended from 1964 back to 1908 using stream gage data from upstream and downstream of the reservoir, as well as reservoir storage data from the old Hogan Dam reservoir.

It should be noted that the time series provided are in fact the total volume of run-off divided by the duration, and are thus average flow values. As a result, it is impossible for an average duration flow value to be greater than a shorter average duration flow. For example, the 30-day flow must by definition be less than or equal to the 15-day flow. Realistically speaking, the 30-day flow will be less than the 15-day flow, because precipitation floods do not begin instantaneously, maintain a constant flow, then instantaneously cease. In most instances the same event accounted for the maximum floods at all durations.

### *Section 1.5: Notable Basins*

The flood records for two high elevation sites, the Kern River at the Isabella Dam (study basin 38) and the Kaweah River at the Terminus Dam (study basin 36), appeared to exhibit a lower bound on flows. Such a lower bound causes the skew coefficient to become very positive, which can in turn cause the fitted LP3 distribution to underestimate the magnitude of large flood quantiles. An in-depth discussion of

this phenomenon can be found in Section 2.1. Through discussion with USACE and USGS, it was determined that this lower bound was likely due to the large snowpack these basins receive every year, which melts slowly over the course of the year, essentially providing a base flow. The typical base flow magnitude was determined graphically by observing the probability plots of the various durations. Ultimately, base flow magnitudes of 150 cfs and 60 cfs were selected for the Kern River and Kaweah River basins respectively. By subtracting these values from the annual maximum rainfall flood series, a more reasonable sample skew coefficient can be estimated. When performing flood frequency analysis for these sites, it would be advisable to subtract the baseflow when fitting the LP3 distribution, and then to add the baseflow to the estimated quantile magnitude. Figures 1.5 and 1.6 plot the 30-day record for the Kern River (study basin 38) with and without baseflow subtraction respectively, and the accompanying fitted LP3 curve. Note that the unadjusted observations appear to exhibit a lower bound and the sample has a highly positive skew coefficient (0.613). USGS and USACE hydrologists felt this was unrealistic. After the baseflow subtraction, the skew coefficient is nearly zero (0.080), which is more comparable to other study basins with similar mean elevation.

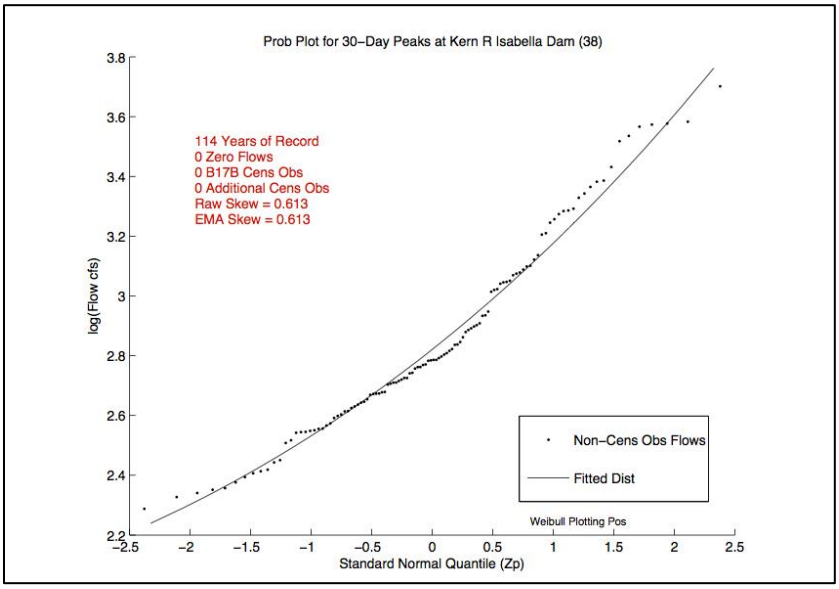


**Figure 1.5:** Probability Plot for Kern River at Isabella Dam (Study Basin 38), without baseflow subtraction.

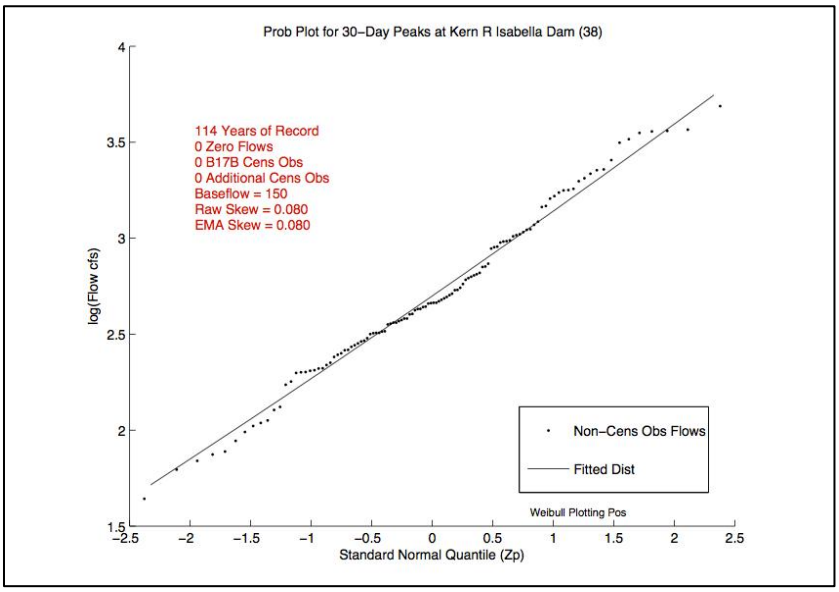


**Figure 1.6:** Probability Plot for Kern River at Isabella Dam (Study Basin 38), with baseflow subtraction.

# REFERENCES


American Society of Civil Engineers, 2006, ASCE California Infrastructure Report Card 2006, California Infrastructure Report Card Committee, ASCE Region 9. Available at *http://www.ascecareportcard.org/*

Brutsaert, W. (2005), Hydrology: An Introduction, Cambridge University Press, New York, NY., pp. 509-550.

Calzascia, E.R., Fitzpatrick, J.A. (1989). "Hydrologic Analysis within California's Dam Safety Program", ASDSO Western Regional Conference and Dam Safety Workshop, May 1-3, 1989, Sacramento, CA.

Carle, David. (2004). "Introduction to Water in California." University of California Press, Berkeley.

Chow, V.T., D.R. Maidment, and L.W. Mays. (1988), Applied Hydrology, McGraw Hill, New York, NY. pp. 350-370.

England, J.F. Jr. and Cohn, T.A. (2007) Scientific and Practical Considerations Related to Revising Bulletin 17B: The Case for Improved Treatment of Historical Information and Low Outliers, American Society of Civil Engineers, EWRI World Water & Environmental Resources Congress, May 15-19, 2007, Tampa, FL, 9 p.

Feaster, T.D., Gotvald, A.J., and Weaver, J.C., 2009, Magnitude and frequency of rural floods in the southeastern United States, 2006—Volume 3, South Carolina: U.S. Geological Survey Scientific Investigations Report 2009-5156, 226 p.

Gotvald, A.J., Feaster, T.D., and Weaver, J.C., 2009, Magnitude and frequency of rural floods in the southeastern United States, 2006—Volume 1, Georgia: U.S. Geological Survey Scientific Investigations Report 2009–5043, 120 p.

Griffis, V. W. (2006). Flood Frequency Analysis: Bulletin 17, "Regional Information, and Climate Change." Ph.D. thesis, Cornell University.

Griffis, V. W., and Stedinger, J. R. (2007a). "Log-Pearson Type 3 Distribution and Its Application in Flood Frequency Analysis. II: Parameter Estimation." *J. Hydrol. Engineering*, 12 (5), 492–500.

Griffis, V.W., and J. R. Stedinger, "Evolution of Flood Frequency Analysis with Bulletin 17." *J. of Hydrol. Engineering*, Volume 12(3), 283-97, 2007b.

Gruber, A.M. and J.R. Stedinger, (2008), Models of LP3 Regional Skew, Data Selection and Bayesian GLS Regression, Paper 596, World Environmental and Water Resources Congress – Ahupua'a, Babcock, R.W. and R. Watson editors, Honolulu, Hawai'i, May 12-16.

Interagency Advisory Committee on Water Data, 1982, Guidelines for determining flood-flow frequency, Bulletin #17B of the Hydrology Subcommittee, Office of Water Data Coordination: U.S. Geological Survey, Reston Virginia, 183 p. Available at *http://water.usgs.gov/osw/bulletin17b/dl_flow.pdf*

Kjeldsen, T. R., and D. A. Jones (2006), Prediction uncertainty in a median-based index flood method using L moments, Water Resour. Res., 42, W07414, doi:10.1029/2005WR004069.

Kjeldsen, T. R., and D. A. Jones (2007), Estimation of an index flood in the UK, Hydrol. Sci. J., 52, 86– 98, doi:10.1623/hysj.52.1.86.

Kjeldsen, T.R., and D.A. Jones (2009), An exploratory analysis of error components in hydrological regression modeling, *Water Resour. Res.*, 45, W02407, doi:10.1029/2007WR006283.

Lamontagne, J.R., J.R. Stedinger, C. Berenbrock, A.G. Veilleux, J.C. Ferris, and D.L. Knifong, 2012. Development of Regional Skews for Selected Flood Durations for the Central Valley Region, California Based on Data Through Water Year 2008, U.S. Geological Survey Scientific Investigations Report 2012-5130 60 p.

Parrett, C., A. Vellieux, , J. R. Stedinger, N. A. Barth, D. Knifong, and J.C. Ferris, 2011. Regional Skew for California and Flood Frequency for Selected Sites in the Sacramento-San Joaquin River Basin Based on Data through Water Year 2006, U.S. Geological Survey Scientific Investigations Report 2010-5260, 94 p.

Reis, D.S., Jr., Stedinger, J.R., and Martins, E.S., 2005, Bayesian generalized least squares regression with application to the log Pearson type III regional skew estimation: Water Resources Research, 41, W10419, doi:10.1029/2004WR003445.

Stedinger, J. R., and Tasker, G. D., 1985, Regional hydrologic analysis, 1, ordinary, weighted and generalized least squares compared: Water Resources Research ,v. 21, no. 9, p. 1421-1432. [with correction, Water Resources Research, v. 22, no. 5, p. 844, 1986.]

Stedinger, J.R., and V.W. Griffis, Flood Frequency Analysis in the United States: Time to Update (editorial), J. of Hydrology, 13(4), 199-204, April 2008.

Stedinger, J. R., Vogel, R. M., and Foufoula-Georgiou, E. (1993). "Frequency Analysis of Extreme Events", in Handbook of Hydrology, chap. 18, pp. 18.1-18.66, McGraw-Hill Book Co., NY.

Tasker, G .D., and Stedinger, J. R., 1986, Regional skew with weighted LS regression: Journal of Water Resources Planning and Management, ASCE, v.112, no. 2, p. 225–237.

Tasker, G.D., and J.R. Stedinger, 1989. An Operational GLS Model for Hydrologic Regression, Journal of Hydrology, 111(1-4), 361–375.

U.S. Army Corps of Engineers, 2002, Sacramento and San Joaquin River Basins comprehensive study: Technical studies documentation Appendix B - Synthetic hydrology technical documentation, accessed May 10, 2010, at *http://www.compstudy.net/docs/techstudies/app_b_synthetichydrology_001.pdf*

Weaver, J.C., Feaster, T.D., and Gotvald, A.J., (2009), Magnitude and frequency of rural floods in the southeastern United States, 2006—Volume 2, North Carolina: U.S. Geological Survey Scientific Investigations Report 2009–5158, 113 p.

Veilleux, A. G. 2009. "Bayesian GLS Regression for Regionalization of Hydrologic Statistics, Floods and Bulletin 17 Skew." MS thesis, School of Civil and Environmental Engineering, Cornell Univ., Ithaca, N.Y.

# CHAPTER 2

## FLOOD FREQUENCY ANALYSIS

Flood risk can be communicated as flood frequency, which is the frequency with which a flood of magnitude $Q_T$ or greater is expected to occur. The return period or recurrence interval $T$ is the average length of time between events of $Q_T$ magnitude or greater. If one assumes that annual peak floods are independent events, the $T$-year flood is the flood which has a $1/T$ annual exceedance probability (AEP) in any given year [Stedinger et al, 1993; Interagency Advisory Committee on Water Data (IACWD), 1982]. To help minimize the correlation between consecutive years' floods, hydrologists in the United States designate the 'water year' to begin on October 1 and end on September 31.

Analytical flood frequency analysis typically involves the fitting of a statistical distribution to the series of annual peak flows, or some transformation of the annual peak flows. The true distribution of annual floods is likely too complex to be understood or of practical use to the analyst [Stedinger et al., 1993]. Without knowledge of the true distribution's form, the discrete empirical distribution from available flood records can be used to estimate the flood distribution at a site. One downside of this approach is that it implicitly assumes floods greater or less than those already observed cannot occur [Loucks and van Beek, 2005; pg 179]. This can present difficulties if one is interested in events with return periods greater than the period of record (POR) for a basin of interest. Flood records in the United States are typically between 10 and 110 years, while design events might have return periods as long as 1,000 years [Calzascia and Fitzpatrick, 1989]. One way to extend the

information contained in data to such extreme events is to fit a parametric probability distribution to the data. Several distributions are reasonable for annual floods, and a variety of alternatives are used in practice around the world [Stedinger et al., 1993, Chow et al., 1988]. In the United States, the log-Pearson Type III (LP3) is used by all Federal agencies as recommended by Bulletin 17B [IACWD, 1982]. This ensures that flood frequency methods employed across the nation are consistent and reasonably precise.

This chapter provides a discussion of Bulletin 17B flood frequency techniques currently in use in the United States, as well as new methodologies which are coming into use. Section 2.1 explores the properties of the LP3 distribution, the importance of the skew coefficient and the use of regional skew estimators. Section 2.2 is a summary of the Bulletin 17B flood frequency procedure. Section 2.3 describes the expected moments algorithm (EMA), which is expected to be adopted as the fitting procedure for an anticipated Bulletin 17C. Finally Section 2.4 discusses the issue of low outliers in flood records, existing identification procedures, and proposed new procedures which are expected to be adopted in a new Bulletin 17C.

### *Section 2.1 Flood Frequency based on the log-Pearson Type III distribution*

The LP3 is a flexible distribution which has been the standard for estimating flood quantiles for gauged sites in the United States since 1982 [IACWD, 1982]. The Pearson Type III distribution is a variation on the gamma distribution which can take a wide variety of shapes depending on its parameters. This section discusses the LP3 distribution in more detail, illustrates the affects of changes to its parameters, and

discusses the importance of the skew coefficient and the use of a generalized skew coefficient.

***Section 2.1.1 The log-Pearson Type III distribution***

The 3-parameter Pearson Type III (P3) distribution is a variation of the two-parameter gamma distribution obtained by subtracting a constant ξ from random variable $X$. The P3 probability density function (pdf) has the form [Bobée, 1975]:

$$f_X(x) = \frac{|\beta|}{\Gamma(\alpha)}\{\beta[x-\xi]\}^{\alpha-1} exp\{-\beta[x-\xi]\} \quad (2.1)$$

which is defined for $\beta(x-\xi) > 0$ and $\alpha > 0$, where $\Gamma(\alpha)$ is the gamma function, which is defined as:

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} \exp(-t)\, dt \quad (2.2)$$

Here $\alpha$ is a shape parameter and $\beta$ is a scale parameter. For $\beta > 0$, $\xi$ is the lower bound for $X$; and for $\beta < 0$, $\xi$ is the upper bound for $X$.

The distribution's three parameters, $\alpha$, $\beta$, and $\xi$, are related to the mean $\mu_X$, standard deviation $\sigma_X$, and skew coefficient $\gamma_X$ of $X$ as follows

$$\alpha = \frac{4}{\gamma_X^2}$$
$$\beta = \frac{2}{\sigma_X \gamma_X} \quad (2.3)$$
$$\xi = \left(\mu_X - 2\frac{\sigma_X}{\gamma_X}\right)$$

Since $\sigma_X$ is positive, $\beta$ and $\gamma_X$ will have the same sign. Thus, for $\gamma_X > 0$, $\xi$ is the lower bound for $X$; and for $\gamma_X < 0$, $\xi$ is the upper bound for $X$.

Random variable $Q$ is said to be log-Pearson Type III (LP3) distributed if its logarithm is P3 distributed. In this study the LP3 distribution is fit to the base-10 common logarithms of the annual peak rainfall flood flow; i.e. $Q$ is the annual peak rainfall flood flow in cubic feet per second (cfs), and $X = \log(Q)$, where $\log(\cdot)$ signifies the base-10 common logarithm. The LP3 pdf has the form [Bobée, 1975]:

$$f_Q(q) = \frac{|\beta|\log(e)}{q\Gamma(\alpha)}\{\beta[\log(q) - \xi]\}^{\alpha-1} exp\{-\beta[\log(q) - \xi]\} \quad (2.4)$$

which is defined $\beta(\log(q) - \xi) > 0$ and $\alpha > 0$, where $\Gamma(\alpha)$ is the gamma function defined in equation (2.2).

The $r^{\text{th}}$ non-central moment of $Q$ is given by [Bobée, 1975]:

$$E[Q^r] = e^{r\xi \log(e)} \left(\frac{\beta \log(e)}{\beta \log(e) - r}\right)^{\alpha} \quad (2.5)$$

where $\beta > r/\log(e)$. Thus, the first three moments of $Q$ are given by:

$$\begin{aligned} \mu_Q &= e^{\xi \log(e)} \left(\frac{\beta \log(e)}{\beta \log(e) - 1}\right)^{\alpha} \\ \sigma_Q^2 &= e^{2\xi \log(e)} \left[\left(\frac{\beta \log(e)}{\beta \log(e) - 2}\right)^{\alpha} - \left(\frac{\beta \log(e)}{\beta \log(e) - 1}\right)^{2\alpha}\right] \\ \gamma_Q &= \frac{E[Q^3] - 3\mu_Q E[Q^2] + 2\mu_Q^3}{\sigma_Q^3} \end{aligned} \quad (2.6)$$

The LP3 is a flexible distribution and can take many shapes, depending on the values of $\alpha$ and $\beta$. Bobée [1975] and Griffis and Stedinger [2007a] illustrate the various shapes the LP3 can take given different combinations of $\alpha$ and $\beta$. Figure 2.1, which is a similar to figure 2 in Griffis and Stedinger [2007a], illustrates the different shapes the LP3 distribution can take given various combinations of $\alpha$ and $\beta$.

By observing equations 2.3 it is clear that only the log-space skew coefficient ( $\gamma_X$) affects the log-space shape of the LP3 probability density function, as neither the log-space standard deviation ( $\sigma_X$) nor the log-space mean ( $\mu_X$) affect the shape parameter $\alpha$. In real-space, however, the log-space standard deviation does impact the shape of the distribution, as is clear in Figure 2.1 [Griffis and Stedinger, 2007a]. The log-space skew coefficient is particularly important in dictating the real-space shape of the LP3 probability density function (pdf) as its sign dictates whether $\xi$ is an upper or a lower bound on Q.

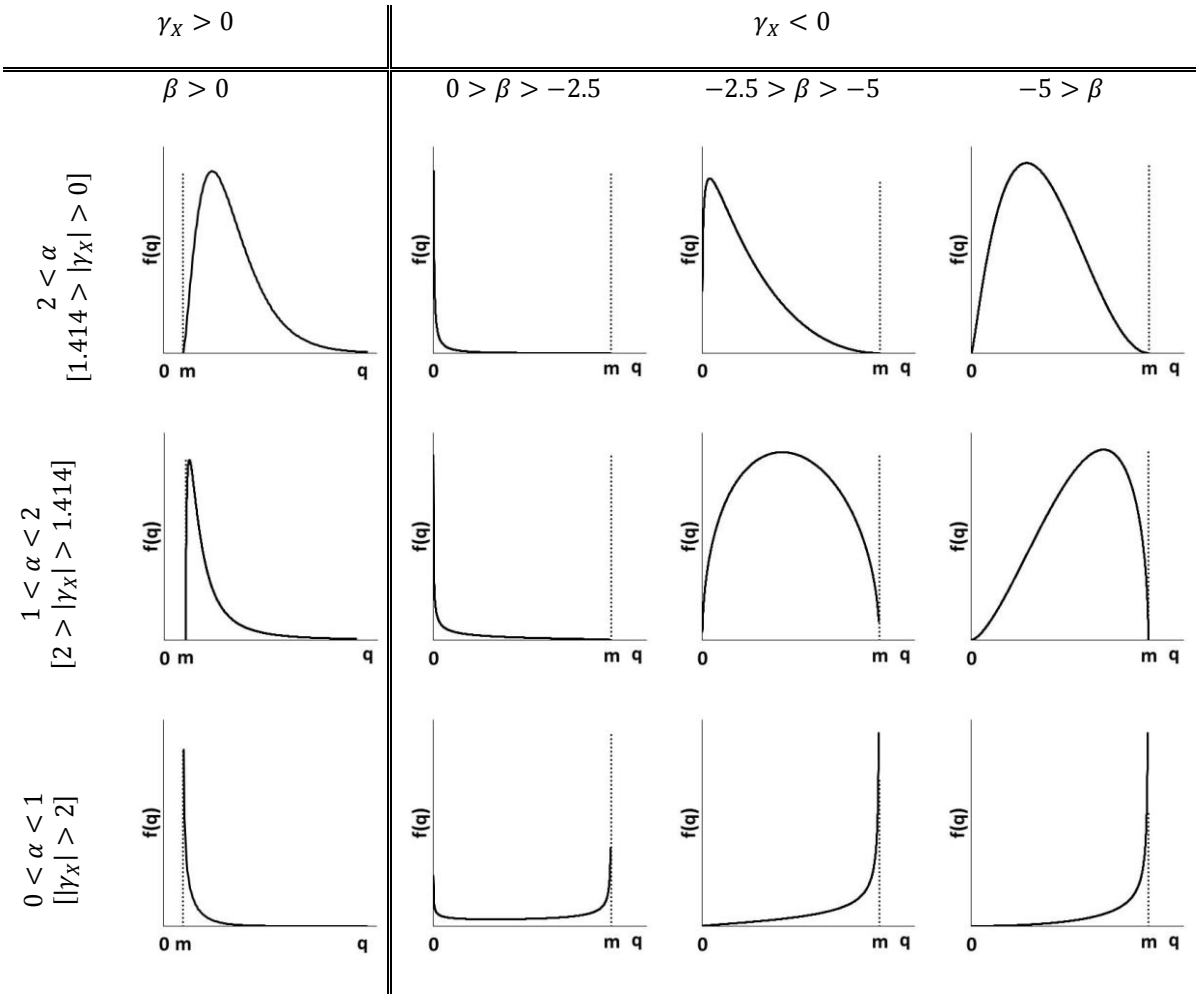


**Figure 2.1:** Probability Density Function for LP3 Distribution with fixed $\boldsymbol{\xi}$ and various combinations of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ (adapted from Griffis and Stedinger 2007a)

Since there are no closed form solutions for the of the LP3, a commonly used alternative formulation to find the $p^{\text{th}}$ quantile, $\hat{Q}_p$, is [IACWD, 1982; Stedinger et al., 1993; Chow et al., 1988]:

$$\hat{Q}_p = 10^{(\mu_X + \sigma_X K_p(\gamma_X))} \tag{2.7}$$

where $K_p(\gamma_X)$ is a frequency factor, which is the $p^{th}$ quantile of a P3 distribution with zero mean, unit variance and skew $\gamma_X$.

The frequency factor $K_p$ can be estimated using tables such as those provided by IACWD [1982] or by an approximation. Bulletin 17B recommends the Wilson-

Hilferty approximation which is valid for estimating frequencies between 0.01 and 0.99 with $\gamma_X \leq |2|$ [Wilson and Hilferty, 1931]:

$$K_p(\gamma) = \frac{2}{\gamma}\left(1 + \frac{\gamma Z_p}{6} - \frac{\gamma^2}{36}\right)^3 - \frac{2}{\gamma} \quad (2.8)$$

where $Z_p$ is the $p^{\text{th}}$ quantile of the standard normal distribution. Kirby[ 1972] provides an alternative approximation which is valid for more extreme skew values. Chowdhury and Stedinger [1991] provide an expression for confidence intervals of LP3 quantile estimates when the population skew is approximated by the sample skew, some regional skew, or a weighted average of the two.

With the formulation provided in equation 2.7, it is not necessary to estimate the LP3 parameters provided in equation 2.3 to estimate the quantiles of the fitted distribution. Bulletin 17B, uses a method-of-moments procedure to estimate flood quantiles by substituting the sample mean, sample standard deviation, and sample skew coefficient into equation 2.7 [IACWD, 1982]. Griffis and Stedinger [2007b] compare this and several other proposed approaches for fitting the LP3. Among these alternative fitting methods the Expected Moments Algorithm (EMA) is most significant to this research, and is discussed in more detail in Section 2.3.

### *Section 2.1.2 Generalized Skew Coefficient*

Use of generalized, or regional, hydrologic variables are commonplace in hydrology. They are particularly useful if one is interested in flood risk assessment at an ungauged location or at a location which has an insufficient record length to ensure the required precision [Griffis and Stedinger, 2007c; Veilleux, 2009; Tasker and Stedinger, 1989]. As an example, the *Flood Estimation Handbook*, which provides recommendations for flood frequency in the United Kingdom, recommends the use of

an index flood, defined as the median annual flood for a basin, which is determined from regional data [Institute of Hydrology, 1999].

As formulated in equation 2.7, the shape of the LP3 distribution is largely determined by the log-space skew coefficient, so precise estimation of the skew coefficient is critical to fitting a data set well. The sample skew coefficient can be difficult to estimate from small samples as it is very sensitive to extreme values [IACWD, 1982]. This is not surprising as the sample skew coefficient estimator cubes each observation's deviation from the mean. In small and moderate sample sizes, even a single extreme observation can have great influence over the sample skew coefficient estimate.

Bulletin 17B advocates the use of a weighted average of the at-site sample skew coefficient and a regional estimate [IACWD, 1982]. This recommendation followed earlier findings by Tasker [1978], who illustrated the utility of a reasonable generalized skew coefficient through Monte Carlo experimentation. In particular, Tasker [1978] was concerned with the weighting factor used to average the at-site sample skew coefficient and the regional skew, and the sensitivity of flood quantile estimates to errors in the regional skew value. Tasker [1978] based his Monte Carlo study on regional skew coefficient estimates with an MSE of 0.302, as this was the recommended MSE for the skew map included in Bulletin 17, the predecessor to Bulletin 17B. Subsequent developments in regional skew estimation by Tasker and Stedinger (1986), Reis et al. (2005), Martins and Stedinger (2002), Gruber et al. (2007), among others, have allowed for regional skew estimates with smaller MSE. Griffis et al. [2004] illustrated the benefits of using more efficient regional skew

values. They found that the precision of quantile estimators is improved through the use of a generalized skew coefficient using Monte Carlo analysis. They generated random samples of sizes between 10 and 100 from Pearson Type III distributions with skew coefficients between -1.0 and 1.0 and fixed mean and standard deviation. Each sample's population skew was randomly generated about a specified regional skew with a specified population skew distribution variance.

The study showed that the benefit of regional skew is greatest for basins with short records, but diminishes as the basin's period of record approaches or surpasses the effective record length of the regional skew model. Generally, the use of a regional skew was shown to decrease the bias of the 99% quantile estimate (the 100-year flood), though it did increase for basins with long record lengths. This was because the period of record exceeded the effective record length of the regional skew coefficient value. The worst of these cases corresponded to only 5% estimation error of the 99% quantile's real space magnitude, so the bias introduced did not represent a significant part of the overall MSE of the estimate [Griffis et al., 2004].

Griffis and Stedinger [2009] repeated the Griffis et al. [2004] Monte Carlo analysis, but considered only a population skew variance of 0.100. They found that use of informative regional skew coefficients decreased the MSE of the 99% quantile estimate, regardless of whether the true variance of the skew coefficient estimate was used or not. They found that when regional skew is equal to zero (in the center of the hydrologic region of interest), there is significant benefit to properly assessing the regional skew precision when estimating the magnitude of the 99$^{\%}$ quantile. When the regional skew coefficient is less than or equal to -0.2, they found no benefit from

properly assessing the estimate precision. In fact, for the case that the regional skew was less than or equal to -0.5, assigning the correct precision to the regional model resulted in less accurate estimates of the 99% quantile. They conclude that use of an informative regional skew coefficient can significantly improve the precision of quantile estimates compared to use of only the sample skew coefficient.

### *Section 2.2 Bulletin 17B Procedure*

Section 1.1: Background of this thesis provides a brief history of the development and motivation for Bulletin 17B, with more detailed discussions found in Griffis and Stedinger [2007b] and Griffis [2006].

This section briefly describes the Bulletin 17B procedure. Other descriptions are provided by Stedinger et al. [1993], Griffis and Stedinger [2007b], and Flynn et al. [2006]. The United States Geological Survey (USGS) also distributes the *PeakFQ* software which automatically performs Bulletin 17B flood frequency analyses on flood records (see Flynn et al. [2006]).

#### *Distribution formulation and skew estimation*

Bulletin 17B recommends fitting the LP3 distribution to the base-10 logarithms of the annual maximum flows to estimate flood quantiles. The properties of the LP3 are discussed in Section 2.1. The logarithm of the $p^{\text{th}}$ flood quantile, or the flood flow associated with an AEP of $p$ can be estimated using the following equation:

$$\log(\hat{Q}_p) = \bar{X} + K_p(G_W)S_X \tag{2.9}$$

where $\bar{X}$ and $S_X$ are the sample mean and standard deviation of the base-10 logarithms of the annual peak flows and $K_p$ is the LP3 frequency factor associated with cumulative exceedance probability $p$ and skew coefficient $G_W$.

Bulletin 17B provides tables for $K_p$ [IACWD, 1982] and also recommends the Wilson-Hilferty transformation (equation 2.8) [Wilson and Hilferty, 1931]. $\bar{X}$ and $S$ are calculated directly using the sample mean and standard deviation formulas respectively, but estimation of the skew coefficient is not as straight forward. As discussed in Section 2.1, the skew coefficient has considerable influence over the shape of the LP3 distribution, and can significantly affect the magnitude of the large flood quantiles which are of most interest. As a result, a good estimate of the skew coefficient is critical to rigorous flood frequency. Because the skew estimator is sensitive to extreme events in small sample sizes, Bulletin 17B recommends using a weighted average of the at-site estimate and a generalized or regional estimate. This weighted average has the form:

$$G_W = \frac{MSE(G_S)G_G + MSE(G_G)G_S}{MSE(G_S) + MSE(G_G)} \tag{2.10}$$

where $G_S$ and $G_G$ are the at-site sample and regional skew coefficients respectively, and $MSE(G_S)$ and $MSE(G_G)$ are the mean square errors for the at-site and regional skew coefficients respectively.

The weights assigned to each skew coefficient estimates in Equation 2.10 depend on the relative magnitude of their MSEs, with the more precise estimate being given more weight. To estimate the MSE of the at-site skew coefficient, Bulletin 17B recommends the following expression:

$$MSE(G_S) = \frac{10^{a+b}}{N^b} \tag{2.11}$$

where $N$ is the years of record and,

$$\begin{aligned} a &= -0.33 - 0.08|G_S| \quad \text{if} \quad |G_S| \leq 0.90 \\ a &= -0.52 + 0.30|G_S| \quad \text{if} \quad |G_S| > 0.90 \\ b &= 0.94 - 0.26|G_S| \quad \text{if} \quad |G_S| \leq 1.50 \end{aligned} \tag{2.12}$$

$b = 0.55$ if $|G_S| > 1.50$

This expression was obtained by fitting a function to Monte Carlo results by Wallis et al. [1974]. Griffis and Stedinger [2009] showed this estimator to be relatively inefficient, yielding errors as high as 10% for $|G_S| \leq 1.414$. Instead they recommend an asymptotically correct expression:

$$MSE(G_S) \approx \left(\frac{6}{N} + a(N)\right)\left(1 + \left\{\frac{9}{6} + b(N)\right\}{\gamma_S}^2 + \left\{\frac{15}{6*8} + c(N)\right\}{\gamma_S}^4\right) \tag{2.13}$$

where $a(N)$, $b(N)$, and $c(N)$ are correction factors for small sample sizes and $\gamma_S$ is the true at-site skew coefficient.

From Monte Carlo studies, Griffis and Stedinger [2007b] recommend:

$$\begin{aligned} a(N) &= \frac{17.75}{N^2} + \frac{50.06}{N^3} \\ b(N) &= \frac{3.93}{N^{0.3}} - \frac{30.97}{N^{0.6}} + \frac{37.1}{N^{0.9}} \\ c(N) &= -\frac{6.16}{N^{0.56}} + \frac{36.83}{N^{1.12}} - \frac{66.9}{N^{1.68}} \end{aligned} \tag{2.14}$$

$MSE(G_S)$ is a function of the record length and the true skew coefficient. As the record length increases, the precision of the skew coefficient estimate increases. $MSE(G_S)$ also increases with the magnitude of the true skew coefficient. Since the true skew coefficient cannot be determined, the sample skew coefficient might be used in equations 2.11 and 2.13. In this case, the estimate of $MSE(G_S)$ will also depend on the sample skew coefficient.

Bulletin 17B recommends three methods for deriving a regional skew coefficient, $G_G$:

> (1) a regional isoline map, developed using a sufficient number of basins which are reasonably close;
> (2) a prediction equation which relates climatologic or basin characteristics to at-site skew;
> (3) the simple arithmetic mean of other regional basin skews can be used.

In the case that no detailed regional skew study has been performed or the analyst prefers not to use such a study, Bulletin 17B provides a national skew map for the United States [IACWD, 1982].

In the case that one of the first three regional skew recommendations are utilized, the analyst must determine an appropriate estimate for $MSE(G_G)$. In the case that the national skew map is used, $MSE(G_G) = 0.302$.

***Outlier identification and treatment***

Bulletin 17B defines outlier observations as 'data points which depart significantly from the trend of the remaining data.' The occurrence of such observations in small samples common to hydrology can undermine the flood frequency analysis. The removal or special treatment of such observations can greatly improve the fit of the LP3 curve and the validity of frequency estimates. Exactly how to classify outliers in practice is not entirely clear and Bulletin 17B cautions that 'all procedures for treating outliers ultimately require judgment involving both mathematical and hydrologic considerations' [IACWD, 1982]. As a guide, Bulletin 17B does recommend the Grubbs-Beck (GB) threshold [Grubbs and Beck, 1972]:

$$T_H = \bar{X} + K_N S$$
$$T_L = \bar{X} - K_N S \quad (2.15)$$

where $T_H$ and $T_L$ are the high and low outlier thresholds respectively. $K_N$ is an LP3 frequency factor corresponding to a one-tailed significance test for the largest (or smallest) magnitude observation given an at-site skew of 0 (corresponding to a log-normal distribution), sample size N, and 10% significance level. Bulletin 17B

provides a table of $K_N$ values for various sample sizes [IACWD, 1982; Appendix 4] and Stedinger et al. [1993] provide the following formula:

$$K_N = -0.9043 + 3.345\sqrt{\log(N)} - 0.4046\log(N) \quad (2.16)$$

If the logarithm of a flood flow is greater than $T_H$ it is considered by Bulletin 17B to be a high outlier. If historical flood data is available which indicates that this flood was the greatest flood in an extended period of time, the flood is treated as historic flood data. Bulletin 17B procedures for the treatment of historic data are described in Appendix 6 of IACWD [1982]. If a flood flow is smaller than $T_L$ it is considered a low outlier and censored from the record. In this case a probability adjustment is recommended. It should be noted that the GB criteria is only a recommendation and that further censoring is often necessary. This point is discussed further in Section 2.4 and in Lamontagne et al. [2013].

***Breaks in the Systematic Record, Observation Thresholds, and Zero Flows***

Systematic flood records often contain breaks (missing years). These breaks can occur for reasons independent of the flood flows in the missing years, such as budgetary or political reasons. These breaks can also be caused by extreme flows in the missing years, for example a large flood might damage gauging equipment. Understanding why a break occurred is important for proper statistical treatment of the record. If the cause of a break is not related to the flood magnitude in the missing year, then the record is treated as a continuous sequence. If breaks in the systematic flood occur due to flood magnitude related events, the record is considered incomplete. In these cases estimates of the missing flood magnitude are often available and can be used [IACWD, 1982].

In the event that flows were below some detection limit, Bulletin 17B provides guidance for a conditional probability adjustment procedure. Conditional probability adjustment is also recommended if a recorded flood magnitude is zero, for which the logarithm is negative infinity. Zero magnitude annual peak floods are not uncommon in arid regions and were experienced in this study.

***Conditional Probability Adjustment***

Given $C$ of $N$ observed flows have been censored, are below a detection limit, or are zeros, an estimator of the probability of an observation being above the censoring threshold is:

$$q_e = \frac{N-C}{N} = 1 - \frac{C}{N} \tag{2.17}$$

Bulletin 17B recommends fitting an LP3 distribution to the $N - C$ retained observations using the procedure described above [IAWCD, 1982]. Let $G(x)$ be the probability density function of this distribution. Flood quantiles greater than the censoring threshold, i.e. flood quantiles whose exceedance probability, $q$ ,is less than $q_e$, can be obtained by solving the following equation for $q$:

$$G(x) = 1 - \frac{q}{q_e} \tag{2.18}$$

This follows because $G(x)$ is really the conditional probability distribution of $X$ given that $X$ is greater than or equal to the truncation level. Thus, the exceedance probability of any recorded flow above the threshold, $q$, is equal to the probability the threshold is exceeded multiplied by probability of the realization $q$ given the probability has been exceeded:

$$q = q_e[1 - G(x)] \tag{2.19}$$

Bulletin 17B offers adjusted product moments based on $G(x)$ which can be combined with the regional skew coefficient for flood frequency analysis. The adjusted product moments are found by using equation 2.18 to estimate the flows whose exceedance probabilities equal 0.50, 0.10, and 0.01: $\hat{Q}_{0.50}$, $\hat{Q}_{0.90}$, and $\hat{Q}_{0.99}$ respectively. Then, the adjusted moments are found using the following formulae [IAWCD, 1982]:

$$G_A = -2.50 + 3.12\left(\log\left(\hat{Q}_{0.99}\right) - \log\left(\hat{Q}_{0.90}\right)\right)/\left(\log\left(\hat{Q}_{0.90}\right) - \log\left(\hat{Q}_{0.50}\right)\right) \quad (2.20)$$

$$S_A = \left(\log\left(\hat{Q}_{0.50}\right) - \log\left(\hat{Q}_{0.50}\right)\right)/(K_{0.01} - K_{0.50}) \quad (2.21)$$

$$\bar{X}_A = \log\left(\hat{Q}_{0.50}\right) - K_{0.50}S_A \quad (2.22)$$

where $K_{0.01}$ and $K_{0.50}$ are the LP3 frequency factors associated with adjusted at-site skew coefficient $G_A$and exceedance probabilities of 0.01 and 0.50 respectively. The approximation given in equation (2.20) is valid for skew coefficients between -2.0 and 2.5 [IAWCD, 1982; Appendix 5]. The adjusted products can be combined with the regional skew coefficient to find flood frequency estimates. This method is recommended for estimating quantiles greater than the median flow (Stedinger et al., 1993). Section 2.3 discusses new and improved approaches for the treatment of low outliers and zero flows when performing flood frequency analysis.

***Section 2.2.1 Regional Skew Coefficient in Bulletin 17B***

As discussed in Section 2.2.1, regional skew coefficient estimation in Bulletin 17B can either involve use of a regional skew study or use of the national skew map [IACWD, 1982; Plate I]. The national skew map is based on initial findings by Hardison [1974] and was added to Bulletin 17 in 1976. The skew map was not changed in the later revisions to Bulletin 17, and is still recommended for national use

and distributed with Bulletin 17B [Hardison, 1974; IAWCD, 1982; Griffis, 2006, Veilleux, 2009].

The national skew map was created using 2,972 basins with drainage areas smaller than 3,000 sq miles and at least 25 years of record [Hardison, 1974; IAWCD, 1982]. The flood records for these sites include data through water year 1973, with no historical data being considered for any site. Basins experiencing serious regulation or diversions which affected the maximum annual flood were removed from the analysis.

No attempt was made to identify or treat high outlier observations. The low outlier procedure described in the earlier Bulletin 17 was used to identify low outliers, which were treated following Bulletin 17 procedures. The primary difference between Bulletin 17B and Bulletin 17 low outlier procedures is the significance level for the Grubbs-Beck threshold in Bulletin 17B has been raised to 10% compared to 1% in Bulletin 17 [IACWD, 1982; p. 12]. This change results in more low outliers being identified, so that regional estimates of skew might be affected [Griffis, 2006]. The developers of the skew map also failed to properly account for zero flows as prescribed by Bulletin 17B. Rather than applying the recommended conditional probability adjustment, the zero flow years were simply omitted from the analysis. Such omissions can greatly impact the at-site estimates of the skew coefficient.

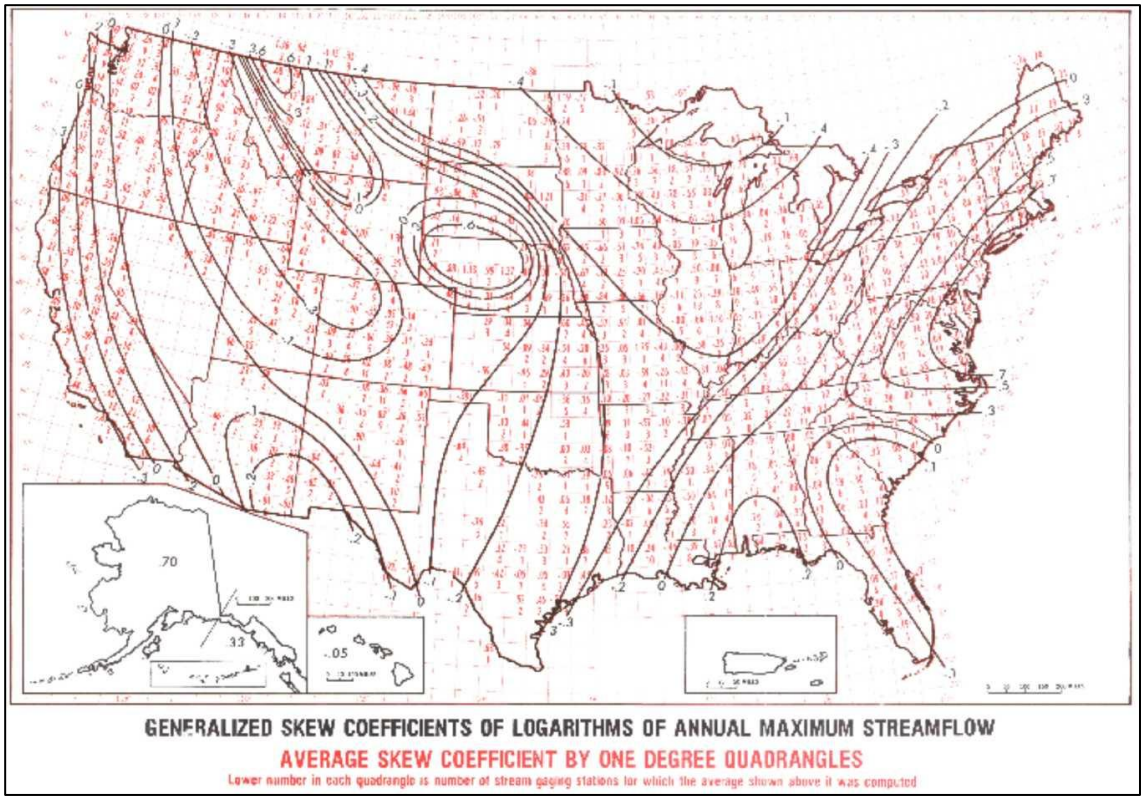


**Figure 2.2**: National Skew Map provided in Bulletin 17B [IAWCD, 1982, Plate I]

Besides the concerns with the development of the skew map cited above, its continued use raises other concerns. The map was developed using flood data through water year 1973 [IAWCD, 1982]. Since then 40 years of additional data have become available which could greatly improve new regional models. Stedinger and Griffis [2008] also note that new and improved statistical methodologies have been developed for regional skew analysis [Reis et al., 2005; Gruber et al., 2007]. When these methods have been applied in various regions of the United States, they have produced MSE of the regional skew coefficient much smaller than the recommended 0.302 for the Bulletin 17B skew map. Recent GLS and WLS/GLS regional skew studies conducted for the Southeast (Veilleux, 2009), California (Parrett et al., 2011), and Iowa (Veilleux et al., 2012) produced effective record lengths of 39 (AVP 0.14) and 55-65 (AVP

0.14), and 50 (AVP 0.13) years respectively compared to the 17 years provided by Bulletin 17B.

Furthermore, it is not entirely clear that all of the skew patterns represented by the isolines on the skew map make hydrologic sense. For example, the Bulletin 17B skew map recommends regional skew values for coastal North Carolina ranging from 0.7 to nearly 0.0 [IACWD, 1982; Plate I]. These skew coefficients span a vast range of LP3 distribution shapes. It is not clear that costal basins in northern North Carolina should exhibit flood characteristics which are so different than those in southern North Carolina. By contrast, the recent Southeast skew study [Veilleux, 2009] produced a regional skew model with constant skew of -0.019 for the entire region.

***Section 2.3 Expected Moments Algorithm***

The Expected Moments Algorithm (EMA) is a moments based method for fitting the LP3 distribution to annual peak flood flows that was first introduced by Cohn et al. [1997]. The EMA was conceived as a more rigorous method to incorporate historical flood data into a frequency analysis than the conditional probability adjustment. Griffis et al. [2004] extend the EMA framework to consider censored observations and use of a regional skew value. In the case that no historical data or low outliers are present in the analysis, the EMA returns identical results to Bulletin 17B. In the case that low outliers are present, the EMA has been shown to perform as well or better than the Bulletin 17B procedures, and is a more theoretically appealing [Griffis et al., 2004].

As roughly half of flood records considered in this study contained at least one low outlier (as identified by GB), it was desirable to utilize the EMA to estimate

sample skew coefficients for each site and duration. A second advantage of utilizing the EMA was that it provides a good estimate of the MSE of the skew coefficient in censored flood records [Cohn et al., 2001]. This estimate was used for flood records which experienced heavy censoring, while the Griffis and Stedinger [2007b] variance formula was used for all other records. In this study four or more censored observations of record length of 70 was considered heavy censoring. This is discussed in more detail in Chapter 4.

A brief description is provided here but longer and more detailed descriptions, with the appropriate equations, can be found in Cohn et al. [1997] and England et al. [2003]. The EMA follows a four step iterative process [England et al., 2003]:

1. Estimate initial sample moments ($\tilde{\mu}_1$, $\tilde{\sigma}_1^2$, $\tilde{\gamma}_1$) from the systematic record; i.e. calculate sample moments based on observed, non-censored flows. $i = 1$.
2. From $i^{\text{th}}$ sample moments, calculate the corresponding LP3 parameters ($\tilde{\alpha}_{i+1}$, $\tilde{\beta}_{i+1}$, $\tilde{\xi}_{i+1}$)
3. Recalculate sample moments ($\tilde{\mu}_{i+1}$, $\tilde{\sigma}_{i+1}^2$, $\tilde{\gamma}_{i+1}$) from LP3 parameters found in step 2 ($\tilde{\alpha}_{i+1}$, $\tilde{\beta}_{i+1}$, $\tilde{\xi}_{i+1}$), but use the entire sample (including censored, zero, and historical flows) using the actual value of the observed but censored flows and the expected value of unobserved flows conditional on the LP3 parameters from the previous iteration ($\tilde{\alpha}_i$, $\tilde{\beta}_i$, $\tilde{\xi}_i$). For $i = 1$ use the censoring or perception threshold for all unobserved flows.
4. Test for convergence of LP3 moments; i.e. check for convergence of ($\tilde{\alpha}_{i+1}$, $\tilde{\beta}_{i+1}$, $\tilde{\xi}_{i+1}$) and ($\tilde{\alpha}_i$, $\tilde{\beta}_i$, $\tilde{\xi}_i$). If convergence is achieved, END. If convergence is not achieved, $i = i + 1$ and return to step 2.

The EMA first estimates sample moments, ($\tilde{\mu}$, $\tilde{\sigma}^2$, $\tilde{\gamma}$), from the non-censored observations. These sample moments are then used to calculate the LP3 parameters, ($\tilde{\alpha}$, $\tilde{\beta}$, $\tilde{\xi}$). Using these LP3 parameters, a new set of sample moments is calculated using all data, which can include observations with known magnitude, magnitude exceeding some threshold, magnitude failing to exceed some threshold, or magnitude

in some interval. In the first iteration of EMA, censored observations and observations with unknown magnitude are represented by the corresponding threshold. In later iterations, these observations are represented by their expected values conditional on the previous iteration's LP3 parameters. This process is repeated until the LP3 parameters between two successive iterations satisfy a convergence criterion.

Implementation of the EMA was done using the 2007 version of *PeakfqSA*, a software package developed and offered by Tim Cohn of USGS. The software can be downloaded for free at:

http://www.timcohn.com/TAC_Software/PeakfqSA/

***Section 2.4 Effects of Low Outliers on Flood Frequency and their Identification***

Recently there has been a movement to revise Bulletin 17B [see Stedinger and Griffis, 2008]. As part of the revision effort, there is renewed interest in outliers in flood records, their interpretation, and tests to identify them.

The Bulletin 17B definition of outliers is 'observations which deviate significantly from the trend in the rest of the data.' While this seems unambiguous, one should note that 'trend in the rest of the data,' might change significantly depending on what one assumes to be 'the rest of the data,' or the *a priori* designated non-outliers. The Bulletin17B definition of an outlier is in fact and intentionally ambiguous. It reminds one of the statement about pornography by Supreme Court Justice Potter Stewart: "I know it when I see it."

In their textbook on outliers, Barnett and Lewis [1994] have a similar, but somewhat more nuanced definition, stating that an outlier is 'an observation (or a subset of observations) which appears to be inconsistent with the remainder of that set

of data,' and that outlier identification must consider how suspected outliers 'appear in relation to the postulated model.' They go on to point out that such observations can 'grossly distort estimates of parameters in the basic model of the population' and can 'frustrate attempts to draw inferences on the original population.'

In flood frequency analysis based on the LP3 distribution, these concerns translate to the inability of a simple parametric probability model to accurately represent the frequency of both large and small floods. Because large floods are of most concern, hydrologist might remove or down-weight the unusually small observations to achieve a better or more robust description of the largest observations.

In the context of Bulletin 17B and fitting the LP3 distribution, the leverage of small observations on the sample skew coefficient and standard deviation, and as a consequence on the fitted distribution is a great concern. Since the skew coefficient is a measure of asymmetry, extreme observations (large or small) can have a great impact on the sample skew, particularly if the record length is short. One expects that extremely large events represent valuable information about the frequency of large floods, so their high influence on the fitted distribution is acceptable. On the other hand, extremely small events often represent different hydrologic processes than large floods, so their high influence on the fitted distribution is disturbing. Klemes [1986] correctly observes "It is by no means hydrologically obvious why the regime of the highest floods should be affected by the regime of flows in years when no floods occur."

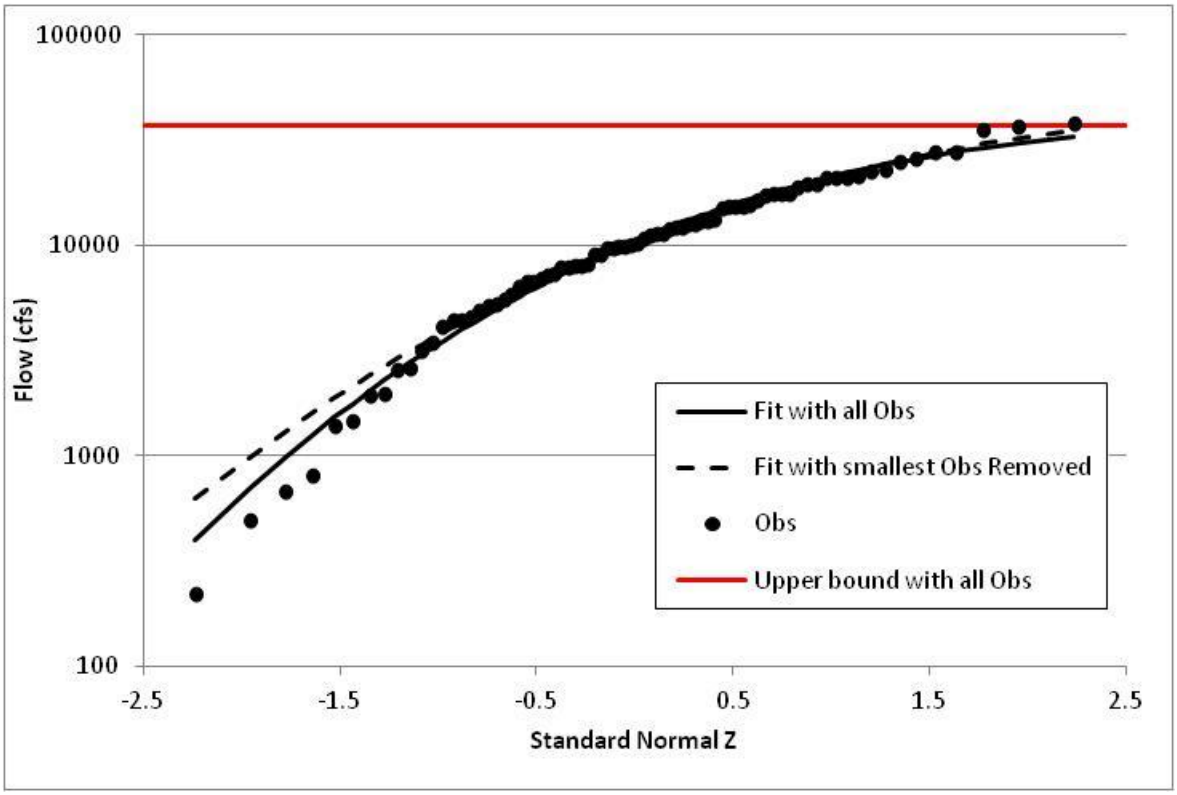


**Figure 2.3**: Probability Plot for 3-day peaks for Putah Creek at Monticello Dam (study site 44), with LP3 fit with all observations and with the smallest observation removed.

As an example, consider the 3-day Putah Creek (study site 44) record, which has a record length of 78 years (see Figure 2.3). When the LP3 parameters are estimated from the entire sample by the method of moments, the result is an upper bound for $X$, $\hat{\xi}$, which was exceeded in 1940. If the smallest flood is simply disregarded, and the distribution is fit with the retained observations, one does not encounter this problem. Is it reasonable to base estimates of extreme flood quantiles on a distribution which claims that observed floods have no chance of occurring? Clearly the smallest observations in this sample are influential low floods, and should be identified and treated in some way to ensure an appropriate fit to the upper tail of the data.

Putah Creek is an extreme example of a common problem: the 3-parameter LP3 is incapable of adequately representing the frequency of both the large and small floods [Spencer and McCuen, 1996; IACWD, 1982; Cohn et al., 2013]. One solution would be to find a better probability model, but US federal agencies are committed to the LP3 distribution. Instead, problematic small values are identified in some way as outliers that are given less weight in the frequency analysis.

Barnett and Lewis [1994] state that outlier identification is ultimately a subjective judgment, still a myriad of tests have been recommended to provide objective guidance for an analyst. These tests can be for a single outlier or for multiple outliers. Tests generally assume some population distribution from which the observations have been drawn. The assumed population distribution then provides a basis for saying an observation is unusual.

The Grubbs-Beck test is recommended by Bulletin 17B for outlier identification in flood samples [IACWD, 1982]. This test is for a single outlier in normal distributed samples. Bulletin 17B provides critical deviates for a 10% test. This means that the test will identify at least one low outlier in 10% of independent random normal samples.

With many samples this test performs well, but in other cases, additional censoring is necessary. Section 2.4.1 describes the Grubbs-Beck test, the visual identification process used in this study, and the new multiple Grubbs-Beck test. Section 2.4.2 compares the three methods when applied to the flood records employed in Chapter 4.

### *Section 2.4.1 Low Outlier Identification Procedures*

*The Grubbs-Beck Test and the Multiple Grubbs-Beck Test*

The Grubbs-Beck low outlier threshold recommended by Bulletin 17B is computed using equation 2.15. Bulletin 17B provides a table of $K_N$ for sample sizes ranging between 10 and 149. This table was taken from Grubbs and Beck [1972], who also provide critical deviates for 0.1%, 0.5%, 1%, 2.5%, 5%, and 10% tests, as well as critical deviates for simultaneously testing the smallest (or largest) two observations in a sample.

Underlying the test Grubbs-Beck test is the statistic,

$$T = \frac{\bar{X} - X_{[1:N]}}{S} \tag{2.23}$$

where $X_{[1:N]}$ is the logarithm of the smallest observation in a sample of size $N$. In a 10% test, $T < K_N$ in 10% of normal samples.

Many flood records contain multiple suspected low outliers, but the Grubbs-Beck test is designed to identify only one outlier in a sample. In the case that multiple suspected low outliers are a concern, it is unclear how the Grubbs-Beck test be applied. One option is to simply apply the test once and identify all observations below the threshold as low outliers (single threshold GB). Another is to apply the Grubbs-Beck test iteratively, removing the outliers identified in each iteration, and then re-computing the threshold on the retained sample (iterated GB).

While nothing is conceptually wrong with the single threshold approach, it rarely identifies more than one outlier, even when subjective judgment would almost certainly identify multiple outliers [Cohn et al., 2013]. The performance of the single threshold GB test is examined with Monte Carlo analysis by Lamontagne et al. [2013].

As expected, the outlier identification rate of the single threshold GB test is 10%, but it rarely identify more than a single outlier (average number of outliers if at least one is identified is 1.012).

Iterated outlier tests, like the iterated GB approach, are common in the outlier test literature [Barnett and Lewis, 1994; Spencer and McCuen, 1996], and is conceptually a very reasonable way to structure a multiple low outlier test. However, when applied to normal samples, the iterated GB identifies multiple low outliers at only a slightly higher rate than the single threshold GB test [Lamontagne et al., 2013]. This is because the Grubbs-Beck critical deviates are based on the distribution of the smallest observation in a normal sample of size *N*. The $k^{th}$ iteration of iterated GB test applies a critical deviate for the $k^{th}$ smallest observation in a sample of size $N-k+1$, to the $k^{th}$ smallest observation in sample of size *N*. This is not the distribution considered in the derivation of the critical values for the GB test. Thus it is not surprising that the test rarely identifies a second outlier in normal samples [Lamontagne et al., 2013].

To address this issue, Rosner [1983] proposed an extreme studentized deviate (ESD) test which is a two-sided generalization of the Grubbs-Beck test. The ESD test simultaneously considers both high and low outliers, and achieves the desired significance level for a pre-specified number of potential outliers. Cohn et al. [2013] propose a single-sided test statistic, similar to the ESD statistic, for any order statistic in a normal sample. The Cohn et al. [2013] test statistic is

$$\widetilde{\omega}_{[k:N]} = \frac{X_{[k:N]} - \hat{\mu}_k}{\hat{\sigma}_k} \tag{2.24}$$

Here $\hat{\mu}_k$ and $\hat{\sigma}_k$ are respectively location and scale parameters based upon $\{X_{[k:N]}, \ldots, X_{[N:N]}\}$ where

$$\hat{\mu}_k = \frac{1}{N-k} \sum_{j=k+1}^{N} X_{[j:N]} \tag{2.25}$$

$$\hat{\sigma}_k = \frac{1}{N-k-1} \sum_{j=k+1}^{N} \left(X_{[j:N]} - \hat{\mu}_k\right)^2 \tag{2.25}$$

The Cohn et al. [2013] test statistic $\widetilde{\omega}_{[k:n]}$ is essentially the same statistic used in the Grubbs-Beck test (Equation 2.23), except that the $k^{\text{th}}$ observation, and all smaller observations have been omitted from the computation of $\widetilde{\omega}_{[k:N]}$. Statistics of this form avoid the problem of masking [Spencer and McCuen, 1996]. Masking is when an outlier causes the test to fail by distorting the test statistic (through the sample mean and standard deviation). The Cohn et al. [2013] statistic avoids this by not including the suspected outliers in the computation of $\widetilde{\omega}_{[k:N]}$.

The major contribution of Cohn et al. [2013] is a quasi-analytical procedure for computing the probability that the test statistic for the $k^{\text{th}}$ observation in a normal sample, $\widetilde{\omega}_{[k:N]}$, is unusually small. Most previous tests were based on critical deviates determined through Monte Carlo analysis, and as a result were limited by the extent of tables provided and the resolution of the original Monte Carlo runs. In fact, the oft cited Grubbs and Beck [1972] is actually just an expansion of the earlier, but limited critical deviates provided in Grubbs [1969]. The advantage of the Cohn et al. [2013] procedure is that it is not limited to a predetermined significance level or

limited to samples less than some maximum *N*, and allows computation of critical value for any $k < N$.

Cohn et al. [2013] do not, however, recommend how their statistic should be applied to test for multiple low outliers. Any application of the statistic would almost certainly be iterative, meaning that various order statistics would be tested in succession. Spencer and McCuen [1996] divide iterative tests into two categories: *forward-step* and *backward-step* tests. *Forward-step*, or outward sweep tests require an *a priori* specification of the maximum number of potential low outliers, $k_{max}$. When the $k^{th}$ observation is tested, and if it is determined to be an outlier, all smaller observations are also identified as low outliers. If the $k^{th}$ observation fails to be identified as a low outlier, the *k-1* observation is tested and so on till no observations remain to be tested, or an outlier is found.

*Backward-step*, or inward sweep tests start by testing if the most extreme (smallest) observation is an outlier. If the smallest observation fails to be an outlier, the test stops. If the smallest observation is identified to be an outlier, the second smallest observation is tested, and so on till an observation fails to be identified as a low outlier. The iterated GB test is a *backward-step* (inward sweep) test. The Rosner [1983] test is a two-sided *forward-step* (outward sweep) test.

A problem that inward sweep tests encounter is *masking*. This occurs when a sample contains multiple low outliers, which affect the test statistic sufficiently (through the mean and standard deviation), that the smallest observation does not look like an outlier. Outward sweeping tests avoid this, difficulty, but *a priori* specification of the number of low outliers can present a challenge in some analyses.

Lamontagne et al. [2013] explores different applications of the Cohn et al. [2013] test to flood records, in anticipation of the test being applied in a proposed *Bulletin 17C*. They refer to iterative applications of the Cohn et al. [2013] test statistic as the Multiple Grubbs-Beck Test (MGBT). Their test consists of three steps. (1) First, starting at the median and sweeping *outward* towards the smallest observation, each observation is tested with a MGBT detection rate, or significance level, of $\alpha_{out}$. If the $k^{th}$ largest observation is identified as a low outlier, the outward sweep stops and all observations less than the $k^{th}$ largest (i.e. i = 1, …, k) are also identified as low outliers. (2) Next an *inward* sweep always starts at the smallest observation and moves towards the median, with a detection rate of $\alpha_{in}$ . If an observation $m \geq 1$ fails to be identified by the inward sweep, the inward sweep stops. The total number of low outliers identified by the MGBT is then the maximum of $k$, and $m - 1$.

Thus, the algorithm has three parameters which must be specified for the three steps:

- 1) Outward Sweep $\alpha$, $\alpha_{out}$
- 2) Inward Sweep from smallest observation $\alpha$, $\alpha_{in}$

After a Monte Carlo analysis, Lamontagne et al. [2013] recommended $\alpha_{out} = 0.005$ and $\alpha_{in} = 0.1$ to be a desirable test configuration for flood frequency guidelines based on the LP3 distribution. The *a priori* maximum number of outliers is specified to *N/2*, thus the initial outward sweep starts at the median. The second step, inward sweep replicates the iterated GB, but uses the correct distribution of the various order statistics. The step 1 outward sweep avoids the *masking* problem by not including any suspected outliers in the computation of the test statistic. The step 2 inward sweep reflects a willingness to identify outliers at a more aggressive rate

($\alpha_{in} = 10\%$), and also a desire to remain consistent with the *Bulletin 17B* 10% GB test.

*Subjective Visual Identification*

The visual identification procedure used on the data discussed in Chapter 4 involved iterative runs of EMA, increasing the low outlier censoring threshold until a satisfactory fit to the largest observations was achieved. This involved first computing the sample moments with the EMA on a record using the GB low outlier identification test. Next, the probability plot of the record was visually inspected. In cases where the fitted distribution failed to describe the frequency of the largest observations, or small observations appeared to deviate significantly from the trend exhibited in the rest of the data, the smallest retained observation was censored and the EMA was re-run. This process was repeated until a satisfactory fit to the largest observations was achieved.

Another consideration for this study was consistency across all flood durations for each site. This study did not consider a single record for each site, as did the previous California instantaneous annual maximum study [Parrett et. al., 2011]. Rather it considered rainfall flood records for five flood durations for each site. In many cases low outlier observations in multiple durations corresponded to the same hydrologic event. Thus, there was a hesitancy to censor an observation in one duration and not in others. In most cases, if an observation appeared as a low outlier in one duration, it also appeared as a low outlier in other durations. In some cases, the flood records did not justify consistent censoring across durations and thus a different number of observations were censored for various durations at the same site. In other

cases, the case for censoring an additional observation for some duration was debatable, but it was a clear low outlier in other durations, so it was censored at all durations for consistency. A table containing the number of low outliers censored for each site and duration can be found in Appendix A of this Thesis.

### *Section 2.4.2 Comparison of Low Outlier Identification Processes*

The purpose of the analysis in Chapter 4 was to construct regional skew coefficient models for rainfall flood durations in California. To this end, it was necessary to estimate the sample skew coefficient for each of the five durations at each of the 50 sites. This was accomplished using the EMA algorithm, with a combination of the single threshold GB and manual identification procedures described above. After this analysis was complete, the MGBT algorithm developed by Cohn et al. [2013] was also applied to the flood records, to assess how its performance compares to subjective manual identification. This section compares the censoring decisions made with the three identification procedures, and provides a few examples of flood frequency outcomes that result from following the recommendations.

This section first reports instances where the single threshold GB test was sufficient, as well as several instances where additional outliers were identified through manual identification. Next, the section compares the performance of the a MGBT configuration with visual identification.

#### *Performance of the Grubbs-Beck test and Subjective Visual Identification*

A list of the number of observations censored for each site and duration for the regional skew analysis can be found in Appendix A of this Thesis. The overall censoring decisions are summarized in Table 2.1. Note that zero flows were

automatically classified as low outliers before applying the GB test or visual identification.

About one third of the 50 sites included in this study have no zero flows or floods treated as low outliers by either the GB test or visual inspection. The 3-day Feather River at Oroville Dam (study basin 13) record displayed in Figure 2.4 is an example.

**Table 2.1:** Summary of low outlier censoring utilized in this study using visual inspection to identify low outliers

| | 1-Day | 3-Day | 7-Day | 15-Day | 30-Day |
|---|---|---|---|---|---|
| **No Censoring** | 16 | 16 | 16 | 16 | 16 |
| **1 Censored** | 28 | 28 | 28 | 28 | 28 |
| **2 Censored** | 1 | 2 | 2 | 2 | 2 |
| **3 Censored** | 1 | 0 | 0 | 0 | 0 |
| **4 Censored**[1] | 2 | 2 | 2 | 2 | 3 |
| **5 Censored**[2,3] | 0 | 1 | 1 | 1 | 0 |
| **>5 Censored**[3] | 2 | 1 | 1 | 1 | 1 |
| **Total** | 50 | 50 | 50 | 50 | 50 |

[1]Cache Creek and N Fork Cache Creek had 4 censored for each duration
[2]Santa Cruz Creek had 6,5,5,5,4 censored observation for 1,3,7,15, and 30-Day durations respectively
[3]Putah Creek had 12, 12, 11, 11, 11 censored observations for 1,3,7,15, and 30-Day durations respectively

The remaining two-thirds of the study sites exhibited at least one flood value that was considered a low outlier at some duration. The GB criterion identified low outliers in about one half of the sites in this study. In many instances, this criterion worked well and resulted in a greatly improved fit.

The 3-day annual peaks for the Trinity River near Coffee Creek (study basin 54) is an example of a site that had a single low outlier (the 1977 peak) identified by the GB test. As shown in Figure 2.5, without censoring, the skew is very negative,

and the largest observations are underestimated. Comparison of Figure 2.5 and Figure 2.6 shows that after censoring the smallest observation, the largest observations are fit much better.

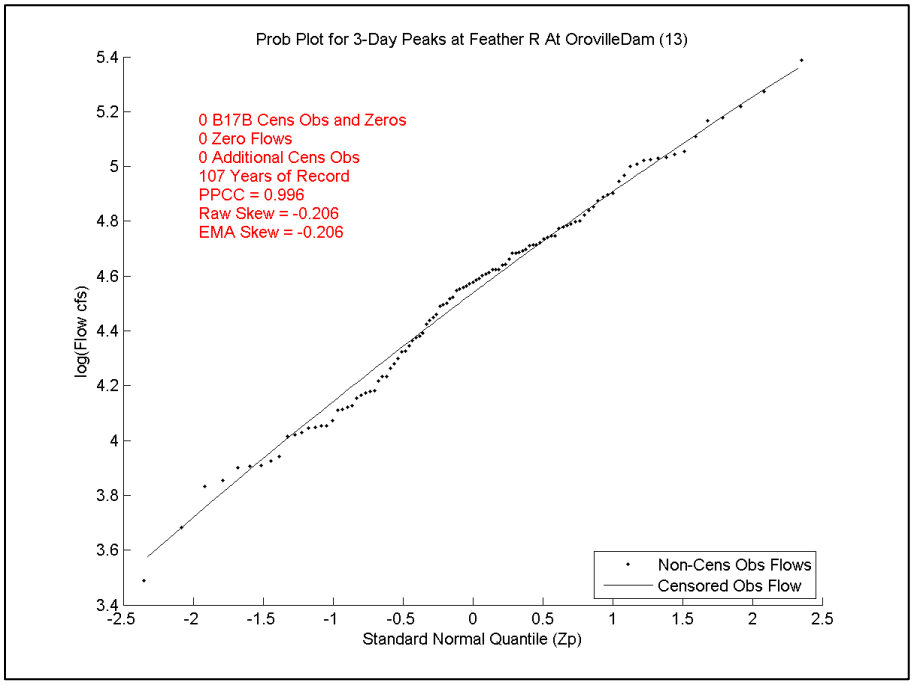


**Figure 2.4**: Probability Plot for 3-day peaks for the Feather River at Oroville Dam.

In this study, the visual identification process typically identified just one or two more observations as low outliers than GB. However, a few sites required more extensive treatment. At these sites, the LP3 distribution appeared to be unable to provide a good fit to both the small and the large observations at the same time. The censoring process at these sites involved iterating EMA, censoring more of the lower tail in order to achieve a good fit in the upper tail. While there was some hesitancy to censor too liberally, high levels of censoring will yield a high MSE of the estimate of the skew coefficient, and in turn a smaller weight in the WLS/GLS regression. This

was deemed a better alternative than dropping the sites completely. Putah Creek at Monticello Dam (study site 44) was the most difficult record in the study. Prior to additional censoring, GB only identified one low outlier, leading to a highly negative skew coefficient. After additional censoring of 11 observations from a record of 51 years, the skew coefficient is still highly negative, -0.741, but the upper tail of the fitted distribution is consistent with the larger observations. In Figure 2.8, the dot depicting the value of the smallest retained observation has 12 times the area of the other points because 12 of the observations in the sample are represented as being less than or equal to that value.

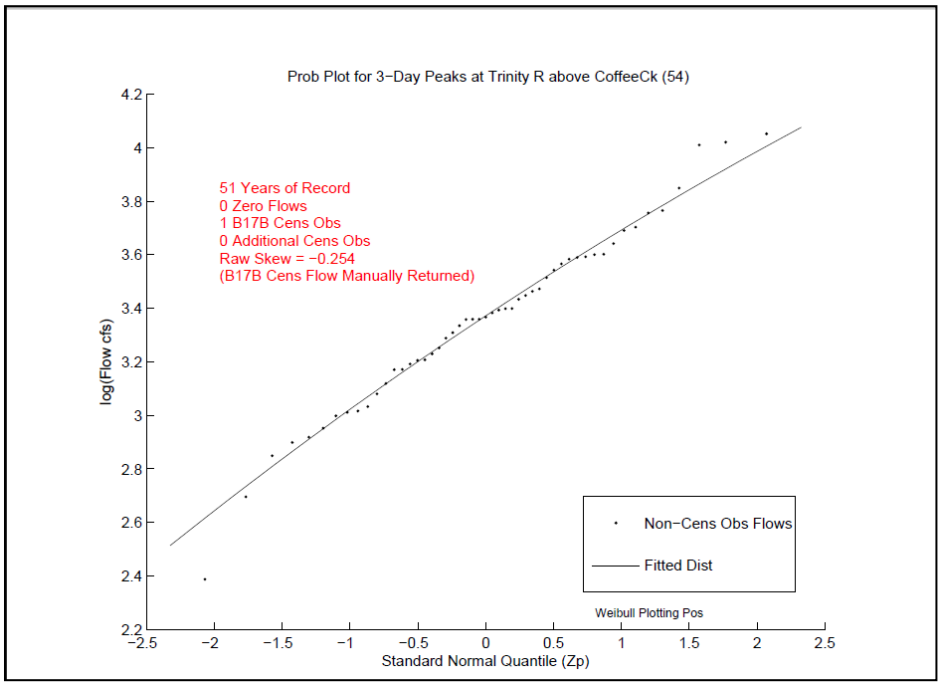


**Figure 2.5**: Probability Plot for 3-day Peaks for the Trinity River at Coffee Creek with no censoring.

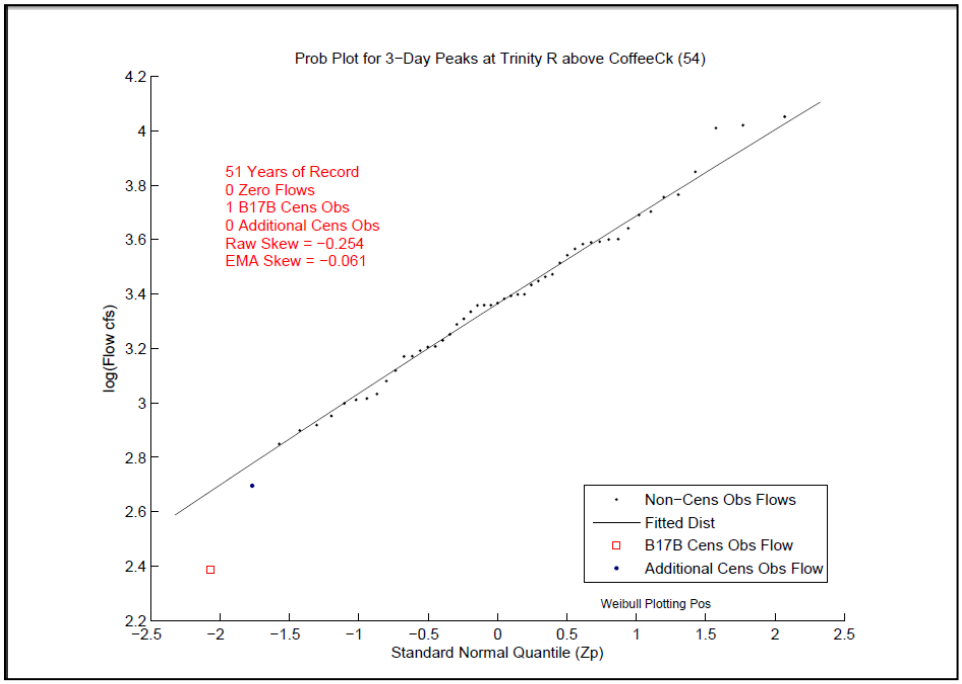


**Figure 2.6:** Probability Plot for 3-Day Peaks for the Trinity River at Coffee Creek with GB censoring.

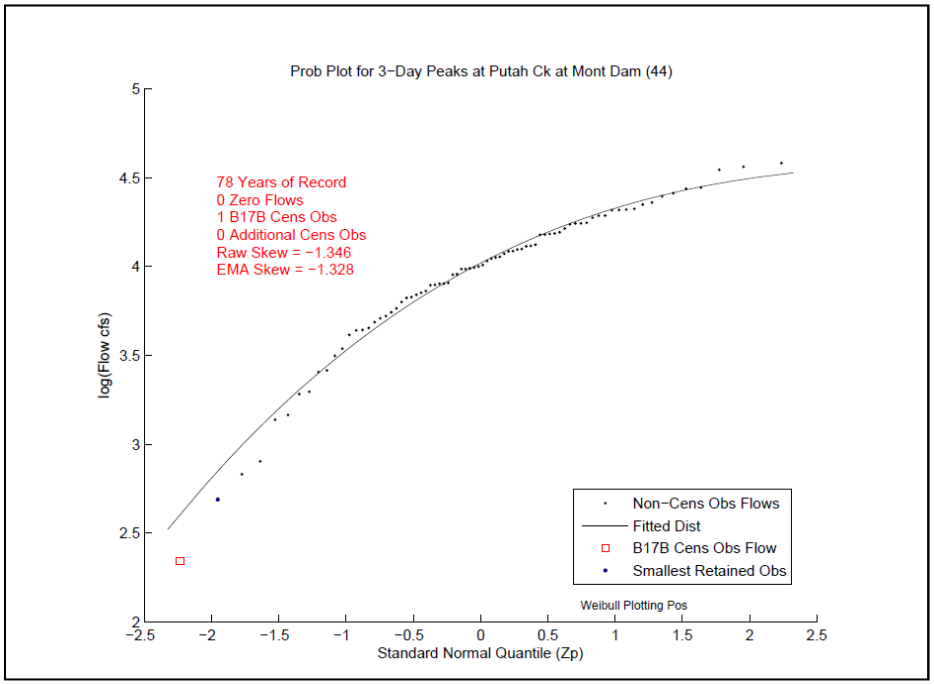


**Figure 2.7:** Probability Plot for 3-Day Peaks for Putah Creek at Monticello Dam with GB censoring of one point.

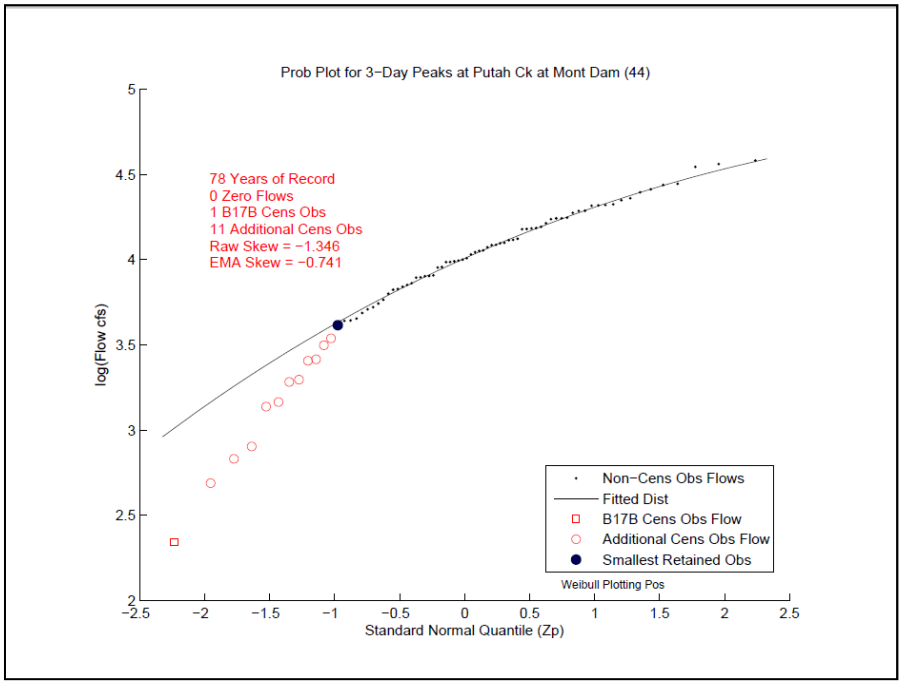


**Figure 2.8:** Probability Plot for 3-Day Peaks for Putah Creek at Monticello Dam after additional censoring

*Performance of the Multiple Grubbs-Beck Test applied to real flood records*

The MGBT was developed to provide the analyst with guidance when encountering difficult records such as Putah Creek (Figure 2.8). Lamontagne et al. [2013] applies the MGBT to samples drawn from different probability distributions to determine a good configuration for flood frequency analysis. In this section, the MGBT configuration with parameters $\{\alpha_{out} = 0.005,\ \alpha_{in} = 0.1\}$ is applied to the rainfall flood duration records. Both the number of outliers and resulting fitted distribution are compared to the results of the visual identification procedure.

Two additional sites which were not included in the analysis in Chapter 4 are also included: Los Banos Creek and Orestimba Creek. These records contained many suspected outliers and zero flows, so while not suitable for the skew study, they are interesting for this section.

Table 2.2 summarizes the number of outliers identified for each duration by single threshold GB, visual identification, and MGBT ($\alpha_{out} = 0.005, \alpha_{in} = 0.1$). Table 2.3 provides a more concise comparison of the low outlier identification procedures as applied to the 50 study sites and the two additional sites.

In no case did the MGBT fail to identify an outlier that was identified by the GB test. This is not surprising because the final inward sweep has an identification rate of 10%, as does the GB test. In about half of all cases the MGBT and GB tests identify the same number of observations. When the MGBT identifies more, it often identifies many more. The GB test identifies more than 3 outliers for only one site (Orestimba Creek, which contains 12 zero flows in a 76 year record). In contrast, The MGBT identifies more than 3 outliers at 17-24 sites, depending on duration, and more than 12 outliers at 11-19 sites, depending on duration.

In about 40-50% of case, depending on duration, the MGBT identifies the same number of outliers as visual identification. Where the two methodologies disagreed, the MGBT identified more low outliers in 75% of cases or more. Most of the cases where visual identification reported more outliers than the MGB were attributable to an attempt to censor consistently across durations. An example of this might be if an observation at the 30-day duration is an outlier and the corresponding observation at the 15-day duration was also censored for no other reason than to maintain consistency. When these cases are accounted for, the MGBT almost always identifies more than the visual identification process. This is an encouraging result: the MGBT is effective at identifying low outliers that trained hydrologists subjectively identified.

**Table 2.2:** Summary of the number records experiencing various levels of low outlier identification by three identification methods, for five durations and 52 sites.

| | 1-Day | | | 3-Day | | | 7-Day | | | 15-Day | | | 30-Day | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Number of Cens Obs | GB | Visual | MGB | GB | Visual | MGB | GB | Visual | MGB | GB | Visual | MGB | GB | Visual | MGB |
| 0 | 25 | 16 | 17 | 27 | 16 | 20 | 25 | 16 | 21 | 23 | 16 | 17 | 23 | 16 | 17 |
| 1 | 24 | 28 | 10 | 22 | 28 | 11 | 24 | 28 | 12 | 26 | 28 | 13 | 26 | 28 | 10 |
| 2 | 1 | 1 | 2 | 1 | 2 | 0 | 1 | 2 | 2 | 1 | 2 | 3 | 1 | 2 | 1 |
| 3 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| 4 | 0 | 2 | 0 | 0 | 2 | 2 | 0 | 2 | 1 | 0 | 2 | 1 | 0 | 3 | 1 |
| 5 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| 6 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| [7,12] | 1 | 1 | 4 | 1 | 1 | 5 | 1 | 1 | 4 | 0 | 1 | 4 | 0 | 1 | 4 |
| [13,20] | 0 | 0 | 3 | 0 | 0 | 3 | 0 | 0 | 2 | 1 | 0 | 3 | 1 | 0 | 3 |
| [21,25] | 0 | 0 | 1 | 0 | 0 | 4 | 0 | 0 | 2 | 0 | 0 | 3 | 0 | 0 | 0 |
| [26,30] | 0 | 0 | 3 | 0 | 0 | 3 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 2 |
| [31,35] | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 1 | 0 | 0 | 2 |
| [35,40] | 0 | 0 | 5 | 0 | 0 | 1 | 0 | 0 | 2 | 0 | 0 | 3 | 0 | 0 | 5 |
| ≥41 | 0 | 0 | 3 | 0 | 0 | 2 | 0 | 0 | 2 | 0 | 0 | 4 | 0 | 0 | 7 |
| Total | 52 | 52 | 52 | 52 | 52 | 52 | 52 | 52 | 52 | 52 | 52 | 52 | 52 | 52 | 52 |

**Table 2.3:** Relative number of outliers identified by the MGBT compared to the GB test and visual identification.

| | 1-Day | 3-Day | 7-Day | 15-Day | 30-Day |
|---|---|---|---|---|---|
| Less than GB | 0 | 0 | 0 | 0 | 0 |
| Same As GB | 26 | 27 | 31 | 22 | 23 |
| More than GB | 26 | 25 | 21 | 30 | 29 |
| Less than Visual | 7 | 6 | 7 | 2 | 5 |
| Same as Visual | 22 | 23 | 26 | 21 | 21 |
| More than Visual | 21 | 21 | 17 | 27 | 24 |

Over identification is generally less of a concern than under identification [Lamontagne et al., 2013]. The logic behind censoring low outliers is that they contain little information about the largest floods, and worse, are potentially exerting undue influence on the magnitude of extreme flood quantiles. In this case, Type II errors are preferred over Type I: by over identifying we are losing observations which contain little information about the frequency of large floods, by under identifying we are retaining observations which potentially distort our probability model. In

operational hydrology, subjective decisions about censoring levels can be difficult to justify. In this study, USGS and USACE hydrologists were often hesitant to subjectively censor more than a few outliers. The MGBT is a useful tool to the analyst in that it provides an objective justification for severe censoring.

*MGBT and Sample Skew Coefficient*

As documented earlier in this section and in Section 4.1, sample log-space skew coefficients for rainfall floods in California are often very negative, which can cause a lack of fit to the largest observations. In the worst case, a highly negative skew can result in an upper bound which is smaller than or nearly equal to observed flows (see Figure 2.3), which is not reasonable. Often, censoring a few of the smallest observations, as recommended by Bulletin 17B results in a better fit to the largest observations, and very often a less negative skew coefficient. One criticism of this practice is that the analyst is simply censoring observations to obtain a skew coefficient which is closer to zero, and thus a flood distribution which is log-normal distributed.

In this study, there were several cases in which additional censoring led to a more negative log-space skew coefficient. One example is the 3-day record for Elder Creek near Paskenta (study site 7). The GB test identified no low outliers, while a visual inspection revealed one low outlier. By censoring this observation, a better fit to the largest observations is achieved, with the skew coefficient changed from -0.864 to -1.007.

The MGBT reveals even more compelling examples where additional censoring actually cause the log-space skew coefficient to become more negative. The

GB test and a visual evaluation of the 30-day record for Salsipuedes Creek near Lompoc (study site 53) identified no low outliers, yielding a log-space skew coefficient of -0.193. MGBT identifies 33 low outliers from a record of 67 years (the most allowed by the algorithm), yielding a log-space skew coefficient of -1.014. Another example is the 30-day record for Big Chico Creek near Chico (study site 10); the GB test and visual inspection identified one low outlier, yielding a log-space skew of -0.584, while the MGB test identified 38 low outliers of a record of 77 years, yielding a log-space skew coefficient of -1.121. Clearly censoring does not always achieve a less negative skew.

In most cases, the additional low outlier identification recommended by the MGBT resulted in a better fit than the GB test and the visual identification procedure. As an example, consider the previously mentioned 3-day record for Putah Creek at Monticello Dam (study site 44) (see Figure 2.7 and Figure 2.8). The GB test identified one low outlier, resulting in a log-space skew coefficient of -1.328, while the visual inspection procedure identified 12 low outliers, resulting in a log-space skew coefficient of -0.741. The MGB test identified 22 low outlier observations, resulting in a log-space skew coefficient of -0.515. Both the visual identification and MGB test fitted distributions appear to fit the largest observations well. A probability plot, with three fitted distributions for the three low outlier identification procedures is plotted in Figure 2.9.

Another example is the 1-day record for Putah Creek. As with the 3-day duration, the GB test identified one low outlier, while the visual identification

procedure identified an additional 11 observations, and the MGB identified 26 low outliers, resulting in a greatly improved fit to the largest observations.

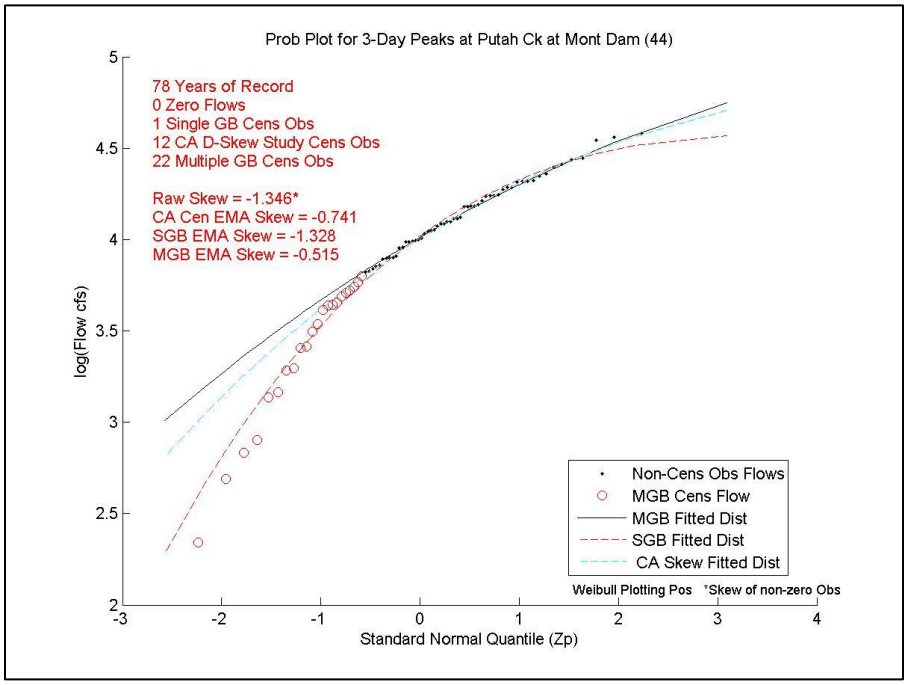


**Figure 2.9**: Probability Plot for 3-day Peaks for Putah Creek at Monticello Dam with the fitted distribution for each of the three low outlier identification procedures.

Another compelling example is the 1-day record for Orestimba Creek near Newman (site not included in skew study). This basin is very arid, experiencing zero flow (no annual rainfall floods) in 12 years of a 77 year record. The GB test identifies no non-zero outliers for this record, resulting in a log-space skew of -1.41, which is the default lower bound for the skew coefficient in the *PeakfqSA* software. A visual inspection of this record was not performed because it was not included in the California duration skew study. Simply disregarding the zero flows and calculating the log-space skew coefficient of the non-zero flows yields a log-space skew of -1.197. By contrast, the MGB test identifies 26 low outliers, yielding a log-space skew of -0.672, which achieves an excellent fit to the largest observations.

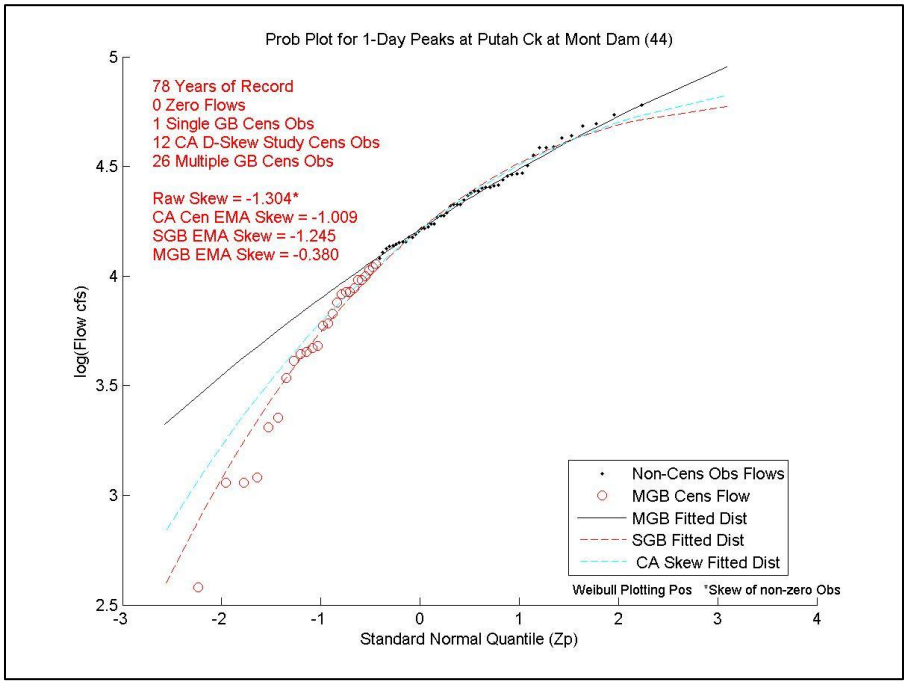


**Figure 2.10:** Probability Plot for 1-day Peaks for Putah Creek at Monticello Dam with the fitted distribution for each of the three low outlier identification procedures.

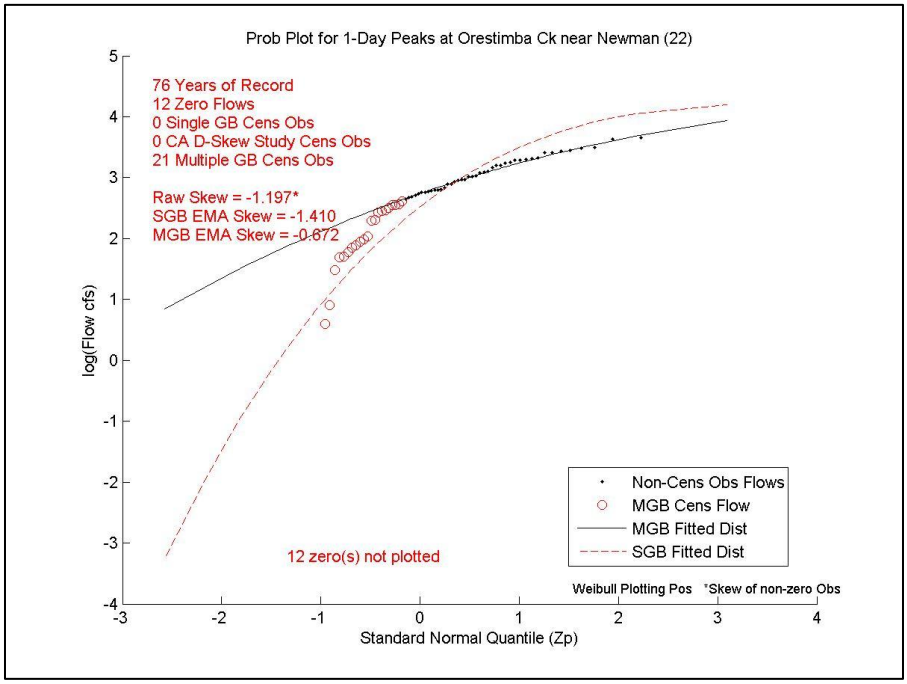


**Figure 2.11:** Probability Plot for 1-Day Peaks for Orestimba Creek near Newman with the fitted distribution for the GB and MGB low outlier identification procedures.

In some cases the additional censoring by the MGB algorithm actually resulted in a worse fit of the largest observations. For example, consider the 1-day records for the Bear River near Wheatland (study site 15) (Figure 2.12) and the Merced River at Exchequer Dam (study site 24) (Figure 2.13). One observation in each of these records was identified as a low outlier either by visual inspection or GB, which provides a reasonably good fit to the largest observations. In both cases, the MGB test identifies all observations smaller than the median as low outliers, which is the most allowed by the MGB algorithm.

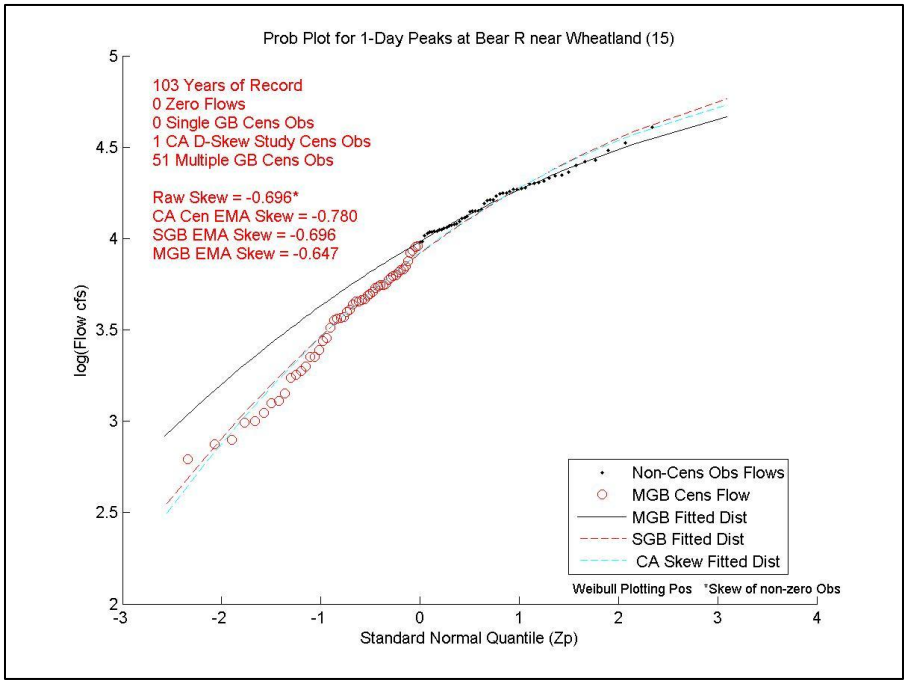


**Figure 2.12**: Probability Plot for 1-Day Peaks for the Bear River near Wheatland with the fitted distribution for each of the three low outlier identification procedures.

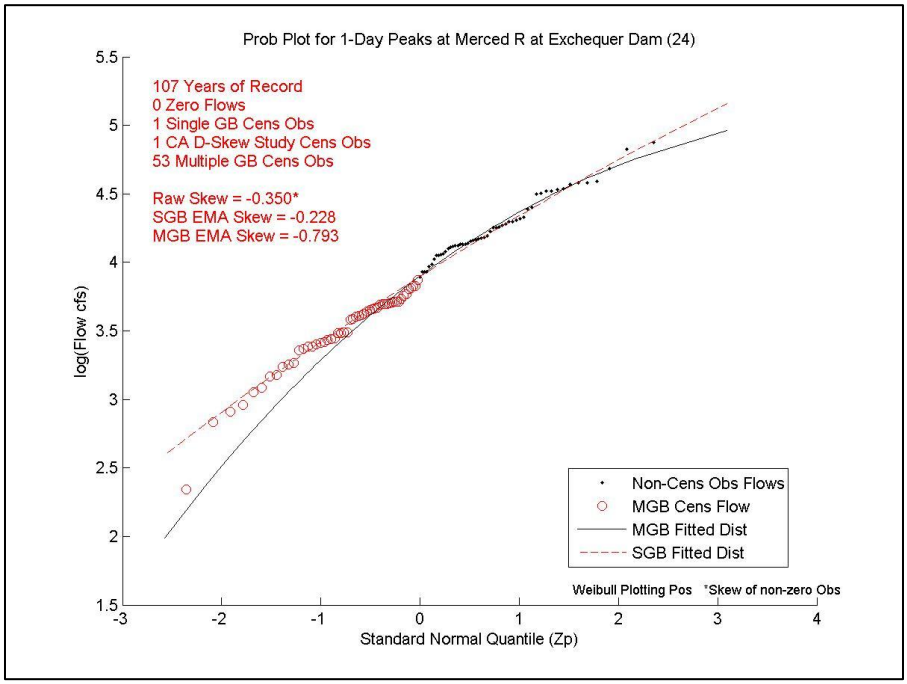


**Figure 2.13:** Probability Plot for 1-Day Peaks for the Merced River at Exchequer Dam with the fitted distribution for the GB and MGB low outlier identification procedure.

The choice of the median as the starting place for the MGB low outlier identification procedure is a logical aspect of the algorithm. The median has a return period of approximately two years, and generally an analyst is only interested in estimating the magnitudes of floods with much smaller exceedance probabilities. However, using the median as the starting place for the MGB procedure is not supported theoretically and is somewhat arbitrary. In practice, the starting place of the MGB procedure should not influence the low outlier identification and the subsequent flood frequency analysis. In this study, when applying the MGB threshold to 52 flood records for five durations (260 flood records in all), it was found that the MGB identified all observations less than the median as low outliers in 12% to 20% of records, depending on duration.

Because the EMA algorithm represents censored flows as smaller than the smallest retained observation, the magnitude of the smallest retained observation has a large impact on the value of the fitted skew coefficient. The arbitrary choice of the median as the starting location for the MGB algorithm appears to be impacting the number of censored observations and by extension the flood frequency results in many cases. Future work will consider alternative starting locations for the MGBT, or alternative algorithms in the case that the censoring threshold is set at the median.

In conclusion, the GB test remains a reasonable low outlier test for flood frequency analysis in many cases, but often fails to identify potential low outliers which cause the fitted distribution to describe the largest observations poorly. The visual inspection procedure used in the California Duration Skew Study often agreed with the GB test, but also often identified additional observations which were classified as low outliers and censored in the flood frequency analysis. In general additional low outlier classification was limited to one or two additional observations, but in an extreme case (Putah Creek, study site 44) involved censoring as many as 12 observations. Though this resulted in better fitting distributions, the process is subjective and different analysts might reasonably censor at different levels. To provide a more objective low outlier identification procedure for arid basins like Putah Creek, the MGBT test has been proposed by Cohn et al. [2013]. This methodology often leads to more liberal censoring than the GB test. In many cases the MGB test results in improved fits, but the algorithm's reliance on the median as the starting location for the low outlier testing appears to be having a deleterious effect in some cases.

One solution for such cases is to initially scan for low outliers starting at the median, but if the median is found to be a low outlier, to scan upwards until a non-outlier observation is found. Another solution might be to scan from the $55^{th}$ or $60^{th}$ percentile if the median is found to be a low outlier. As with the GB test, low outlier identification with MGB should involve a review, as no automated detection algorithm can anticipate all cases.

***Conclusion***

This chapter describes flood frequency analysis procedures which are commonly applied in operational hydrology. Section 2.1 describes the log-Pearson Type III distribution and the use of a generalized skew coefficient when fitting it to short flood records. This motivates the procedure described in subsequent chapters. Section 2.2 summarizes the Bulletin 17B procedure, which is the standard flood frequency procedure followed by Federal agencies in the United States. Section 2.3 describes the Expected Moments Algorithm, which is a moments based fitting procedure for the log-Pearson Type III distribution which can accommodate many different data types, censoring levels, and use of a generalized skew coefficient. Finally, Section 2.4 describes the low outlier identification procedure applied in this study, as well as a new objective outlier detection test which is better suited to the detection of multiple outliers in a flood sample.

REFERENCES


Barnett, V., T. Lewis. (1994). *Outliers in Statistical Data*, John Wiley & Sons, New York.

Bobee, Bernard. (1975)."The Log Pearson Type 3 Distribution and Its Application in Hydrology." Water Resources Research II, no. 5 (October 1975).

Calzascia, E.R., Fitzpatrick, J.A. (1989). "Hydrologic Analysis within California's Dam Safety Program", ASDSO Western Regional Conference and Dam Safety Workshop, May 1-3, 1989, Sacramento, CA.

Chow, V. T., D. R. Maidment, and L. W. Mays. (1988). *Applied Hydrology*, McGraw-Hill, New York.

Chowdhury, J.U., Stedinger, J.R., 1991. Confidence interval for design flood with estimated skew coefficient. Journal of Hydraulic Engineering 117 (7), 811–831.

Cohn, T., W. L. Lane and W. G. Baier, (1997). "An Algorithm for Computing Moments-Based Flood Estimates when Historical Flood Information is Available,"*Water Resources Research*, 33(9), pp. 2089-2096, 1997.

Cohn, T., W. L. Lane and J. R. Stedinger, (2001). "Confidence Intervals for EMA Flood Quantile Estimates," *Water Resources Research*, 37(8), 2001.

Cohn, T. A., J. F. England, C. E. Berenbock, R. R. Mason, J. R. Stedinger, J. R. Lamontagne, (2013). "A generalized Grubbs-Beck test statistic for detecting multiple potential influential low outliers in flood series." *Water Resources Research*, in press.

England, J. F., Salas, J. D. & Jarrett, R. D. (2003), 'Comparisons of two moments-based estimators that utilize historical and paleoflood data for the log Pearson type III distribution.', Water Resources Research 39(9), 1243.

Flynn, K. M., Kirby, W. H. & Hummel, P. (2006), 'Users Manual for PeakFQ, Annual Flood Frequency Analysis Using Bulletin 17B Guidelines: U.S. Geological Survey Techniques and Methods Chapter 4 of Book 4, Section B'.

Griffis, V. W. (2006). "Flood Frequency Analysis: Bulletin 17, Regional Information, and Climate Change." Ph.D. thesis, Cornell University.

Griffis, V. W. , J.R. Stedinger, and T. A. Cohn . (2004). "LP3 Quantile Estimators with Regional Skew Information and Low Outlier Adjustments", *Water Resources Research, 40*, W07503, doi:1029/2003WR002697.

Griffis, V. W., and Stedinger, J. R. (2007a). "Log-Pearson Type 3 Distribution and Its Application in Flood Frequency Analysis. I: Distribution characteristics." *J. Hydrol. Engineering*., 12 (5), 482–491.

Griffis, V. W., and Stedinger, J. R. (2007b). "Log-Pearson Type 3 Distribution and Its Application in Flood Frequency Analysis. II: Parameter Estimation." *J. Hydrol. Engineering*, 12 (5), 492–500.

Griffis, V. W., and J. R. Stedinger, (2007c), The Use of GLS Regression in Regional Hydrologic Analyses, *J. of Hydrology*, 344(1-2), 82-95, [doi:10.1016/j.jhydrol.2007.06.023].

Griffis, V.W., and J. R. Stedinger. (2009). "Log-Pearson Type 3 Distribution and Its Application in Flood Frequency Analysis. III: Sample Skew and Weighted Skew Estimators", *J. of Hydrol. Engineering* 14(2), pp. 121-130.

Gruber, A.M., D.S. Reis Jr., and J. R. Stedinger. (2007). "Models of regional skew based on Bayesian GLS regression", *World Environmental & Water Resources Conference-Restoring out Natural Habitat*, edited by K.C. Kabbes, Tampa, Florida. May 15-18, Paper 40927-3285.

Grubbs, F.E. (1969). "Procedures for Detecting Outlying Observations in Samples." *Technometrics*, Vol. 11, No. 1:1-21

Grubbs, Frank E., and Glenn Beck. (1972). "Extension of Sample Sizes and Percentage Points for Significance Tests of Outlying Observations." *Technometrics* 14, no. 4: 847-854.

Hardison, C.H. (1974)."Generalized Skew Coefficents of Annual Floods in the United States and Their Application." *Water Resources Research* 10, no. 5: 745-752.

Interagency Advisory Committee on Water Data. (1982). Guidelines for determining flood-flow frequency, Bulletin #17B of the Hydrology Subcommittee, Office of Water Data Coordination: U.S. Geological Survey, Reston Virginia, 183 p. Available at *http://water.usgs.gov/osw/bulletin17b/dl_flow.pdf*

Institute of Hydrology. (1999), Flood Estimation Handbook, Wallingford, U.K.

Kirby, W., 1972. Computer oriented Wilson–Hilferty transformation that preserves the first 3 moments and lower bound of the Pearson type 3 distribution. Water Resources Research 10 (2), 220–222.

Klemes, V.K. (1986), "Dilettantism in hydrology: Transition or Destiny?", *Water Resources Research*, 22(9S), 177S-188S.

Lamontagne, J.R., J.R. Stedinger, T.A. Cohn, and N. Barth, Robust National Flood Frequency Guidelines: What is an outlier?, in *World Environmental & Water Resources Conference, Cincinnati OH*, edited by C.L. Patterson, S.D. Struck, and D.J. Murray, ASCE EWRI, 2013.

Loucks, D., van Beek, E. Water Resource Systems Planning and Management. (2005). Paris: United Nations Educational, Scientific, and Cultural Organization.

Martins, E. S., and Stedinger, J. R. (2002). "Cross-correlation among estimators of shape," Water Resources Research, v. 38, no. 11, 1252, doi: 10.1029/2002WR001589

Parrett, C., A. Vellieux, , J. R. Stedinger, N. A. Barth, D. Knifong, , and J.C. Ferris, 2011. "Regional Skew for California and Flood Frequency for Selected Sites in the Sacramento-San Joaquin River Basin Based on Data through Water Year 2006", OFR, U.S. Geological Survey.

Reis, D.S., Jr., Stedinger, J.R., and Martins, E.S. (2005). "Bayesian generalized least squares regression with application to the log Pearson type III regional skew estimation," Water Resources Research, 41, W10419, doi: 10.1029/2004WR003445.

Rosner, B. (1983), Percentage Points for a Generalized ESD Many-Outlier Procedure, Technometrics, 25(2), pp. 165-172.

Spencer, C.S. and McCuen, R.H. (1996). "Detection of Outliers in Pearson Type III Data." J. of Hydro. Eng. 1(1).

Stedinger, J. R., Vogel, R. M., and Foufoula-Georgiou, E. (1993). “Frequency Analysis of Extreme Events”, in Handbook of Hydrology, chap. 18, pp. 18.1-18.66, McGraw-Hill Book Co., NY.

Stedinger, J.R., V.W. Griffis. (2008). “Flood Frequency Analysis in the United States: Time to Update.” *J. Hydrol. Engineering*. 199-204.

Tasker, G.D. (1978). “Flood Frequency Analysis with a Generalized Skew Coefficient,” Water Resources Research, 14 (2), 373-376.

Tasker, G .D., and Stedinger, J. R., (1986), “Regional skew with weighted LS regression.” *Journal of Water Resources Planning and Management*, ASCE, v.112, no. 2, p. 225–237.

Tasker, G.D., and J.R. Stedinger. (1989). An Operational GLS Model for Hydrologic Regression, Journal of Hydrology, 111(1-4), 361–375.

Veilleux, A. G. (2009). “Bayesian GLS Regression for Regionalization of Hydrologic Statistics, Floods, and Bulletin 17 Skew.” Ph.D. thesis, Cornell University.

Veilleux, A.G., J.R. Stedinger, and D. A. Eash, Bayesian WLS/GLS Regression for Regional Skewness Analysis for Crest Stage Gage Networks, Paper 227, Crossing Boundaries, Proceedings World Environmental and Water Resources Congress, ed. Eric D. Loucks, Amer. Society of Civil Engineering, Albuquerque, New Mexico, pp. 2253-63, May 20-24, 2012.

Wallis, J. R., N.C. Matalas, and J.R. Slack. (1974)."Just a Moment." *Water Resources Research* 10, no. 2: 211-219.

Wilson, E. B. and M.M. Hilferty.(1931). “The Distribution of Chi-Square,” *Proc., National Academy of Science*, Vol. 17, No 12, December 1931, pp. 684-688.

# CHAPTER 3

# LEAST SQUARES REGRESSION METHODS FOR REGIONAL SKEW MODELS

Regionalization of hydrologic variables is commonplace in hydrology [Stedinger and Tasker, 1985; Reis et al., 2005;Institute of Hydrology, 2005; Griffis and Stedinger, 2007; Chow et al., 1988; Brutsaert, 2005]. Regional models might be constructed to provide estimate a variable of interest at a site with no data. In other cases a regional model might be constructed to aid the estimation of a noisy statistic from limited data, as in the use of a regional skew value for flood frequency analysis in Bulletin 17B [IACWD, 1982]. Estimation of the skew coefficient is difficult in small samples, as it is very sensitive to extreme observations. In Bulletin 17B, the weighted average of the regional skew and the at-site sample skew is used in subsequent analyses, where the weights depend on the relative magnitude of the MSE of each estimate. The smaller the MSE of the regional model, the more meaningful it will be for Bulletin 17B flood frequency analysis. Bulletin 17B provides a national skew map, which was originally published in 1974 [Hardison, 1974], and has a MSE of 0.302. The result is that the regional skew is of little value to the overall analysis, even for relatively short records.

This chapter describes regression methods which are used to create regional skew models with much smaller MSE, and are therefore more meaningful to flood frequency analysis. Section 3.1 describes the historical development of weighted and generalized least squares methods for regional skew models. Section 3.2 presents the theoretical derivation of these methods and a new hybrid method developed for this

study. Section 3.3 describes the concept of redundant basins, their statistical implications under the GLS framework, and metrics for their detection. Section 3.3 also discusses recent attempts to explicitly account for model error correlation in a GLS framework.

***Section 3.1: Development of WLS/GLS skew coefficient models***

Tasker and Stedinger [1986] proposed a Weighted Least Squares (WLS) procedure for regressing regional skew models on basin characteristics, given sample skew values for a set of sites. This procedure weights sample skew coefficients on the basis of sampling error (a function of record length) and model error variance describing the precision of the model. Stedinger and Tasker [1985, 1986a,b] and Tasker and Stedinger [1989] laid out a Generalized Least Squares (GLS) procedure for quantile regression. The main advantage of GLS over a WLS regional skew regression is that GLS explicitly accounts for cross-correlation among skew coefficient estimators. This is an important consideration, because highly cross-correlated samples are not independent samples. Failure to account for this cross-correlation can lead to overestimation of model precision [Stedinger, 1983; Stedinger and Tasker, 1985]. Reis et al. [2005] provide a Bayesian analysis of that GLS model. Bayesian GLS is an improvement over traditional GLS because it generates the posterior distribution of the model error variance. Traditional GLS can return a zero model error variance estimate, indicating that no model error exists, which is unreasonable [Reis et al. , 2005].

Bayesian GLS methods have been applied to regional skew studies for annual peak flows in the Illinois River basin [Reis et al., 2005] and the Southeast [Feaster et al., 2009; Gotvald et al., 2009; Weaver et al., 2009].

Recently, a modified WLS/GLS procedure was applied to a regional skew study for annual peak flows in California [Parrett et al., 2011]. It was found that skew coefficients in California exhibit much higher cross-correlations than those in the Southeast. As a result, standard GLS became unstable and produced complex weights which were not justifiable [Veilleux, 2009]. This modified methodology has been applied in Iowa [Eash et al., 2013], Arizona [Mason, 2012], Missouri [Mason, 2012], and Vermont, with more studies underway [Veilleux, 2013 personal communication].

### *Section 3.2: Standard GLS and Hybrid WLS/GLS Procedure*

This section describes the theoretical development of the regional regression framework applied to regional skew regression in Chapter 4. This section also describes the theoretical development of various diagnostic statistics which are adapted to the described regression framework.

#### *Section 3.2.1 Standard GLS Framework*

The GLS framework assumes that a linear model with additive errors can adequately describe the regional skew coefficient [Reis et al., 2005]. If one has $k$ explanatory variables and $n$ study basins, the GLS model has the form:

$$\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_k x_{ik} + \varepsilon_i \qquad \text{for } i = 1, \ldots, n \tag{3.1}$$

In more compact vector-matrix notation, the GLS regional model in equation (3.1) can be written:

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{3.2}$$

where

$\widehat{\boldsymbol{Y}}$ is an $(n \times 1)$ vector of unbiased at-site skews for each study basin, $\hat{y}_i$
$\boldsymbol{X}$ is an $(n \times k)$ vector of basin characteristics for each study basin
$\boldsymbol{\beta}$ is a $(k \times 1)$ vector of model coefficients
$\boldsymbol{\varepsilon}$ is an $(n \times 1)$ vector of total error.

Following the regression framework from Tasker and Stedinger [1986], the total error vector $\boldsymbol{\varepsilon}$ has two components: the regression model error and the sampling error of the at-site sample skew coefficient estimates for each site. The regression model error, $\delta$, is due to the use of an imperfect model. The sampling error is due to lack of data and is a function of record length and the true at-site skew coefficient. The total error vector $\boldsymbol{\varepsilon}$ by assumption has zero mean (i.e. $E[\boldsymbol{\varepsilon}] = \mathbf{0}$), and covariance matrix, $\boldsymbol{\Lambda} = E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T]$.

Let $\sigma_\delta^2$ be the variance of the regression model error and $\sum(\hat{\boldsymbol{y}})$ be the covariance matrix of the at-site sample skew coefficients for the $n$ sites. The $i^{\text{th}}$ diagonal element of $\sum(\hat{\boldsymbol{y}})$ is the variance of the at-site sample skew coefficient for site $i$. The covariance matrix $\boldsymbol{\Lambda}$ of the total error vector $\boldsymbol{\varepsilon}$ for the model with an additive error described in Equation (3.2) is given by [Tasker and Stedinger, 1986]:

$$\boldsymbol{\Lambda} = \sigma_\delta^2 \boldsymbol{I} + \Sigma\left(\widehat{\boldsymbol{Y}}\right) \tag{3.3}$$

In a WLS regression, the covariance of the sampling errors are ignored and the off-diagonal terms of $\boldsymbol{\Lambda}$ are zero. In a GLS regression, the off-diagonal terms of $\boldsymbol{\Lambda}$ represent the covariance of the sampling errors of the at-site sample skew coefficients [Stedinger and Tasker; 1985]. Given the covariance matrix $\boldsymbol{\Lambda}$, the unbiased minimum variance estimator of $\boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}}$, is [Draper and Smith, 1967; equation 2.11.10]:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{\Lambda}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{\Lambda}^{-1} \widehat{\boldsymbol{Y}} \tag{3.4}$$

Because $\boldsymbol{\Lambda}$ is not known, it must be estimated from the data [Kroll and Stedinger, 1998, 1999; Reis et al., 2005]. This is important, particularly when assessing model uncertainty, because basins with high covariance do not represent independent samples. It is expected that basins that are spatially close will experience similar hydrologic conditions, leading to high covariance among estimated flood characteristics. To correctly analyze regional skew model uncertainty, a reasonable covariance model must be developed. Griffis and Stedinger [2009] provide an expression for the variance of the at-site sample skew coefficient $\hat{\gamma}_i$ (Equation (3.22)) but no direct expression for estimation of the covariance of the at-site sample skew coefficients exists. Martins and Stedinger [2002] provide an empirical relationship between the correlation of the skew coefficient and the annual flood peaks between two sites. Thus an important step in regional skew studies is the development of an appropriate regression model for the cross-correlation of annual floods at different basins. This is discussed further in Section 3.2.3.

A variety of regression diagnostic statistics have been developed for GLS to help in model comparison and selection. A brief summary and description of some useful statistics is provided below. See also Gruber et al. [2007].

***Variance of Prediction and Average Variance of Prediction***

Variance of prediction is the variance of the model's predicted skew $\tilde{y}_i$ about the true value $y_i$. Reis et al. [2005] make the distinction between old sites, which were included in the original skew study, and new sites which were not. For old sites already included in the study, the model error variance and the sampling variance of

the predicted skew are correlated. For a new site not included in the construction of the regional model, this correlation does not exist [Veilleux, 2009]. The variance of prediction for a new site is given by [Reis et al., 2005]:

$$VP_{new}(i) = E\left[\sigma_\delta^2\right] + \boldsymbol{x}_i(\boldsymbol{X}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{X})^{-1}\boldsymbol{x}_i^T \tag{3.5}$$

where $\boldsymbol{x}_i$ is the $i^{th}$ row of $\boldsymbol{X}$, containing the basin characteristics of site $i$.

$VP_{new}$ can be thought of as the sum of the model error variance and the sampling variance of the predicted skew for site $i$. The variance of prediction for an old site is given by [Reis et al., 2005]:

$$VP_{old}(i) = E\left[\sigma_\delta^2\right] + \boldsymbol{x}_i(\boldsymbol{X}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{X})^{-1}\boldsymbol{x}_i^T - 2E\left[\sigma_\delta^2\right]\boldsymbol{x}_i(\boldsymbol{X}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Lambda}^{-1}e_i \tag{3.6}$$

where $e_i$ is a column vector with one on the $i^{\text{th}}$ row and zero otherwise.

The third term in the expression for $VP_{old}(i)$ accounts for the correlation of the model error variance and the sampling error for site $i$ which was included in the construction of the regional skew model. For sites not included in the original study, this term is zero, thus the expression for variance of prediction at a new site in equation (3.5).

When comparing various regional models, the models' performance over a region is likely of more interest than for a specific site. Tasker and Stedinger [1986] recommend an average variance of prediction statistic, which is obtained by averaging the sampling error over the basins used to generate the regional model:

$$AVP_{new} = \frac{1}{n}\sum_{i=1}^{n} VP_{new}(i) \tag{3.7}$$

Gruber et al. [2007] notes that $AVP_{new}$ as formulated in equation (3.7) assumes basins used in the construction of the regional model are characteristic of those where

the regional skew will be applied. A natural extension of $AVP_{new}$ to old sites is provided by Reis et al. [2005]:

$$AVP_{old} = \frac{1}{n}\sum_{i=1}^{n} VP_{old}\,(i) \tag{3.8}$$

Again, recall that $VP_{old}$ includes a term for the cross correlation of the model error variance and sampling error variance for a site included in the study. This term is not included in the $VP_{new}$ because a new site is not included in the study.

***Pseudo $R_\delta^2$***

The standard $R^2$ statistic commonly reported in regression studies is the based on the ratio of the residual errors and the total variability observed in the data. In the Tasker and Stedinger [1985] regression framework the errors are divided into model errors due to an imperfect model and sampling errors due to finite record lengths. Thus, even a perfect model cannot explain all of the variability observed in the data, but instead will have zero model error variance. Thus, the traditional $R^2$ is not appropriate for the Tasker and Stedinger [1985] regression framework.

Griffis and Stedinger [2007] develop a pseudo $R_\delta^2$ statistic which is a measure of model performance. The pseudo $R_\delta^2$ can be calculated as:

$$R_\delta^2 = 1 - \frac{\hat{\sigma}_\delta^2(k)}{\hat{\sigma}_\delta^2(0)} \tag{3.9}$$

where

$\hat{\sigma}_\delta^2(k)$ is the model error variance of a model with $k$ explanatory variables
$\hat{\sigma}_\delta^2(0)$ is the model error variance of a model with no explanatory variables.

Thus, a regression model with no model error would have an $R_\delta^2$ of one, while a model which performs no better than the constant model would have an $R_\delta^2$ of zero. This is a

measure of the amount of variation in the true skew the model is describing. The constant model captures no variation in the skew, so it must always have an $R_\delta^2$ of zero.

***Average sampling error variance***

The average sampling error variance (ASEV) describes the contribution of the sampling error of regional model parameters to the total variance of prediction [Stedinger and Tasker, 1985]. The ASEV plus the expected value of the model error variance is $AVP_{new}$ as defined in equation (3.7). ASEV can be calculated as:

$$ASEV = \frac{1}{n}\sum_{i=1}^{n}\left[\boldsymbol{x}_i(\boldsymbol{X}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{X})^{-1}\boldsymbol{x}_i^T\right] \tag{3.10}$$

***Error Variance Ratio and Misrepresentation of Beta Variance***

The error variance ratio (EVR) is a measure of the relative magnitudes of the sampling error and model error. The motivation for calculation of EVR is to determine if an OLS analysis is appropriate or if a more sophisticated WLS or GLS analysis may be necessary. The EVR can be calculated as [Griffis and Stedinger, 2007]:

$$EVR = \frac{\sum_{i=1}^{n} var(\hat{y}_i)}{n\hat{\sigma}_\delta^2} \tag{3.11}$$

OLS considers only a single error term, whereas WLS and GLS divide the error into sampling error and model error. If sampling error is negligible compared to model error, an OLS analysis should be sufficient. If sampling error is important, then a WLS or GLS analysis is preferred to an OLS analysis. Griffis and Stedinger [2007] advise that 20% is a reasonable threshold above which WLS or GLS should be

employed. Another issue to consider is whether $var(\hat{y}_i)$ varies widely from site-to-site, suggesting that the homoscedasticity assumption of OLS is less applicable.

Misrepresentation of beta variance is used to determine if a WLS analysis is sufficient or if GLS is preferred [Griffis and Stedinger, 2007]. GLS considers the cross-correlation of the at-site sample skew coefficients, while WLS neglects this cross-correlation. Stedinger and Tasker [1985] show this cross-correlation has the greatest impact on the precision of the estimator of the constant term. If the cross-correlations are insignificant, then the WLS estimate of the variance of the constant term of the regional WLS model should be nearly the same as the GLS estimate of the variance. If the cross-correlations are positive and significant, WLS would overestimate the precision of the constant term, so that the WLS estimate of the variance would be significantly smaller than the GLS estimate of the variance. Veilleux et al. [2011] define MBV as:

$$MBV = \frac{Var\left[\hat{\beta}_0^{WLS} \middle| GLS\ Anlaysis\right]}{Var\left[\hat{\beta}_0^{WLS} \middle| WLS\ Anlaysis\right]} = \frac{\boldsymbol{w}^T \boldsymbol{\Lambda} \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{v}} \tag{3.12}$$

where $\boldsymbol{v}$ is a $n \times 1$ vector of ones and

$$w_i = \frac{1}{\Lambda_{ii}} \tag{3.13}$$

If WLS is appropriate, MBV should not be much greater than one. Griffis and Stedinger [2007] advise that to avoid standard error of the constant term greater than 10%, a threshold of 1.2 should be adopted.

***Leverage and Influence***

Leverage is a measure of an observation's potential to influence the model regression due to its location in the variable space. Tasker and Stedinger [1989] show

that leverage for site $i$ in a GLS framework is given by the $i^{\text{th}}$ diagonal element of the $\boldsymbol{H}^{*}_{GLS}$:

$$\boldsymbol{H}^{*}_{GLS} = \boldsymbol{X}\{\boldsymbol{X^T\Lambda^{-1}X}\}^{-1}\boldsymbol{X^T\Lambda^{-1}} \tag{3.14}$$

The influence of an observation is the effect removing that observation from the analysis has on the final model. Observations whose omission causes great change in the analysis are said to be influential, or to have high influence [Weisberg, 1985]. The influence for an observation in an OLS regression is computed:

$$D_i = \frac{\hat{\varepsilon}_i^2 h_{ii}}{k(1-h_{ii})^2\hat{\sigma}^2} \tag{3.15}$$

where

$h_{ii}$ is the $i^{\text{th}}$ diagonal element of the OLS hat matrix, $\boldsymbol{H}_{OLS}$ (defined in equation (3.16))
$\hat{\varepsilon}_i$ is the residual for observation $i$
$\hat{\sigma}^2$ is the observed variance of the residuals.

The OLS hat matrix, $\boldsymbol{H}_{OLS}$, is:

$$\boldsymbol{H}_{OLS} = \boldsymbol{X^T}\{\boldsymbol{X^TX}\}^{-1}\boldsymbol{X} \tag{3.16}$$

Tasker and Stedinger [1989] extended Cook's D to the generalized least squares case. Noting that $h_{ii}/(1-h_{ii}) = Var\,(\hat{y})/Var(\hat{\varepsilon}_i)$, Cook's D for the GLS case becomes:

$$D_{GLS,i} = \frac{h^{*}_{GLS,ii}\hat{\varepsilon}_i^2}{k\left(\lambda_{ii}-h^{*}_{ii}\right)^2} \tag{3.17}$$

where

$\lambda_{ii}$ is the $i^{\text{th}}$ diagonal element of $\boldsymbol{\Lambda}$
$h^{*}_{ii}$ is the $i^{\text{th}}$ diagonal element of $\boldsymbol{H'} = \boldsymbol{H}^{*}_{GLS}\boldsymbol{\Lambda}$

Leverage and influence values in excess of $2k/n$ and $4/n$ respectively are considered large by Tasker and Stedinger [1989].

***Effective Record Length***

Effective record length (ERL) is often cited when comparing various regional skew models (see Griffis and Stedinger 2007; Veilleux, 2009; Gruber et al., 2007; Reis et al, 2005). ERL is obtained by substituting $VP_{new}(i)$ and $\tilde{y}_i$ for $var(\hat{\gamma})$ and $\hat{\gamma}$ in Equation (3.22) and solving for the resulting $n$. Thus, ERL is a function of the model predicted skew and its variance. The ERL for a specific site may not be informative as an indicator of the overall model performance if the model contains non-constant terms. In this case an average ERL might be reported, in which the $AVP_{new}$ and $\hat{\beta}_0$ from the constant model are used. This approach is taken by Lamontagne et al. [2012].

***Section 3.2.2 Hybrid WLS/GLS Procedure***

Bayesian GLS as formulated in Section 3.2.1 has been used in regional skew studies in the Illinois River Basin [Reis et al., 2005] and the Southeast [Veilleux, 2009; Weaver et al., 2009; Feaster et al., 2009; and Gotvald et al., 2009]. Parrett et al. [2011] document that when cross-correlations of the annual maximum series are high, as with California annual maximum flow series, the GLS regression produces extreme weights (both positive and negative). If the true correlation relationships were known these weights would defensible. Because sample correlations and correlation models are used, these complex weights are not defensible. To address this problem, Parrett et al. [2011] developed a WLS/GLS hybrid method which used WLS to find the model coefficients and GLS to estimate the precision of the models. Because WLS ignores the high cross-correlations between sampling errors, it produced reasonable weights but would generally underestimate the precision of the model. Thus, a GLS analysis

which acknowledges the cross-correlations was used to estimate the precision of the estimated parameters. This WLS/GLS procedure was used in regional skew analysis for rainfall floods in California [Chapter 4, Lamontagne et al., 2012] and in Iowa [Eash et al., 2013]. Ongoing studies are currently applying the methodology to Arizona [Mason, 2012], Missouri [Mason, 2012], and Vermont, with more studies planned [Veilleux, 2013 personal communication].

The degree of cross-correlation among the annual maximum rainfall floods were even greater in this study than in the annual instantaneous maximum study. This is likely because only rainfall floods were considered whereas Parrett et al. [2011] considered floods of every origin. Also, averaging the flood volume over the flood duration dampened the effect of spatial and temporal variability among hydrologic events. Because of the high cross-correlations in this study, a similar methodology to that of Parrett et al. [2011] was utilized. This section discusses the procedure for this method as well as adjustments to various diagnostic statistics.

***Step 1: Ordinary Least Squares Analysis***

The first step of the hybrid analysis is to obtain an OLS skew coefficient estimate for each study basin. This is an iterative procedure. At first a simple constant model, which is an average skew value, is used. After the subsequent WLS and GLS estimators are computed, the OLS model can be expanded to reflect those basin characteristics which are shown to be statistically significant. This was shown to improve the average variance of prediction of the final model, but did not generally affect which basin characteristics were statistically significant.

The OLS model has the form described in Equation (3.2). Recall that $\widehat{\boldsymbol{Y}}$ is a vector of unbiased skew coefficients. The unbiased skew coefficient for site $i$, $\hat{y}_i$, can be found using the correction factor derived by Tasker and Stedinger [1986], given below

$$\hat{y}_i = \left[1 + \frac{6}{N_i}\right] G_i \tag{3.18}$$

where

$N_i$ is the length of systematic record at site $i$, excluding any historical record
$G_i$ is the traditional sample skew coefficient estimator

The OLS regression parameters, $\widehat{\boldsymbol{\beta}}_{OLS}$, are simply (Drapper and Smith, 1967; equation 2.1.17):

$$\widehat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\widehat{\boldsymbol{Y}} \tag{3.19}$$

The OLS regional skew coefficient vector, $\widetilde{\boldsymbol{y}}_{OLS}$, is given by

$$\widetilde{\boldsymbol{Y}}_{OLS} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{OLS} \tag{3.20}$$

The OLS regional skew coefficient for site $i$ is the $i^{\text{th}}$ element of $\widetilde{\boldsymbol{Y}}_{OLS}$, notated: $\tilde{y}_{OLS,i}$. Tasker and Stedinger [1986] provide an expression for the variance of the unbiased skew coefficient estimator based upon:

$$Var[\hat{y}_i] = \left[1 + \frac{6}{N_i}\right]^2 Var[G_i] \tag{3.21}$$

Griffis and Stedinger [2009] provide an equation for the variance of the at-site sample skew coefficient. When combined with the equation (3.21), the variance of the at-site unbiased skew is calculated as:

$$\boldsymbol{\Sigma}(\widehat{\boldsymbol{Y}})_{ii} = Var[\hat{y}_i]$$

$$= \left(1 + \frac{6}{N_i}\right)^2 \left(\frac{6}{N_i + a(N_i)}\right)\left(1 + \left(\frac{9}{6} + b(N_i)\right)\left(\tilde{y}_{OLS,i}\right)^2 + \left(\frac{15}{48} + c(N_i)\right)\left(\tilde{y}_{OLS,i}\right)^4\right) \tag{3.22}$$

where

$$a(N_i) = -\frac{17.75}{N_i^2} + \frac{50.06}{N_i^3}$$

$$b(N_i) = \frac{3.92}{N_i^{0.3}} - \frac{31.1}{N_i^{0.6}} + \frac{34.86}{N_i^{0.9}} \tag{3.23}$$

$$c(N_i) = -\frac{7.31}{N_i^{0.59}} + \frac{45.9}{N_i^{1.18}} - \frac{86.5}{N_i^{1.77}}$$

$a(N_i)$, $b(N_i)$, and $c(N_i)$ in equation (3.23) are correction factors for small sample sizes. The variance of the at-site skew coefficient estimate calculated in equation (3.22) is generally used in the following WLS and GLS steps. The effects of minor censoring (less than 3 censored observations in 70 years of record) on the variance estimate in equation (3.22)(3.17) were generally considered insignificant and were ignored. For sites experiencing heavier censoring of low outliers and zero flows, the variance of the skew coefficient generated by EMA was substituted in. This is examined in more detail in Chapter 4.

***Step 2: Weighted Least Squares analysis***

WLS is used to develop estimators for the regression coefficients. Unlike OLS, WLS explicitly accounts for heteroscedacity of the at-site sample skew coefficient estimates. Unlike GLS, WLS neglects the correlation between study basins.

The first step in the WLS analysis is estimation of the model error variance using Bayesian-WLS (B-WLS) as described by Reis et al. [2005].

The B-WLS estimate of the model error variance, $\hat{\sigma}^2_{\delta,B-WLS}$, is the mean of the posterior distribution of $\sigma^2_\delta$ from a WLS analysis, $E\left[\sigma^2_{\delta,B-WLS}\right]$. Reis et al. [2005] layout a quasi-analytical procedure to numerically estimate the pdf of the marginal distribution of $\sigma^2_\delta$, which is given as

$$f\left(\sigma^2_{\delta,B-WLS}|\widehat{\boldsymbol{Y}}\right) \propto \left[\left|\boldsymbol{\Lambda}_{WLS}\left(\sigma^2_{\delta,B-WLS}\right)\right| \left|\boldsymbol{X}^T\boldsymbol{\Lambda}_{\mathrm{WLS}}\left(\sigma^2_{\delta,B-WLS}\right)^{-1}\boldsymbol{X}\right|\right]^{-0.5} \cdot \exp\left[-0.5\left(\widehat{\boldsymbol{Y}}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{WLS}\right)^T\boldsymbol{\Lambda}_{\mathrm{WLS}}\left(\sigma^2_{\delta,B-WLS}\right)^{-1}\left(\widehat{\boldsymbol{Y}}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{WLS}\right)\right]\xi\left(\sigma^2_{\delta,B-WLS}\right) \tag{3.24}$$

where

$\boldsymbol{\Lambda}_{\mathrm{WLS}}\left(\hat{\sigma}^2_{\delta,B-WLS}\right)$ is the WLS covariance matrix defined in equation (3.27)
$\widehat{\boldsymbol{\beta}}_{WLS}$ are the WLS model parameters
$\xi(\sigma^2_{\delta,B-WLS})$ is the prior distribution of $\sigma^2_{\delta,B-WLS}$
$|\boldsymbol{A}|$ denotes the determinant of matrix $\boldsymbol{A}$.

The MEV prior distribution, $\xi(\sigma^2_\delta)$, is assumed to be exponential, having the form

$$\xi\left(\sigma^2_{\delta,B-WLS}\right) = \lambda e^{-\lambda\left(\sigma^2_{\delta,B-WLS}\right)}, \qquad \sigma^2_{\delta,B-WLS} > 0 \tag{3.25}$$

A value of 10 was assigned to $\lambda$,which corresponds to a prior mean model error variance of 1/10. This indicates that that $P\left(\sigma^2_{\delta,B-WLS} \leq 0.3\right) = 0.95$. The Bayesian estimate of $\hat{\sigma}^2_{\delta,B-WLS}$ can be calculated as:

$$\hat{\sigma}^2_{\delta,B-WLS} = E\left[\sigma^2_{\delta,B-WLS}|\hat{\boldsymbol{y}}\right] = \int \sigma^2_{\delta,B-WLS} f\left(\sigma^2_{\delta,B-WLS}|\hat{\boldsymbol{y}}\right) d\sigma^2_{\delta,B-WLS} \tag{3.26}$$

This is implemented using numerical integration.

Given $\hat{\sigma}^2_{\delta,B-WLS}$, the covariance matrix, $\boldsymbol{\Lambda}_{WLS}$, is a diagonal matrix calculated with equation (3.27), with off-diagonal elements set to zero.

$$\boldsymbol{\Lambda}_{WLS}\left(\hat{\sigma}^2_{\delta,B-WLS}\right) = \hat{\sigma}^2_{\delta,B-WLS}\boldsymbol{I} + diag\left(\Sigma(\widehat{\boldsymbol{Y}})\right) \tag{3.27}$$

where

$\boldsymbol{I}$ is an $(n \times n)$ identity matrix ($n$ is the number of study basins)
$diag\left(\sum\left(\widehat{\boldsymbol{Y}}\right)\right)$ is an $(n \times n)$ matrix containing the variances of unbiased at-site skew coefficient estimates obtain through equation (3.22) on the diagonals and zero otherwise

The WLS weight matrix, $\boldsymbol{W}$ is computed:

$$\boldsymbol{W} = \left[\boldsymbol{X}^T\boldsymbol{\Lambda}_{\text{WLS}}\left(\hat{\sigma}^2_{\delta,B-WLS}\right)^{-1}\boldsymbol{X}\right]^{-1}\boldsymbol{X}^T\boldsymbol{\Lambda}_{\text{WLS}}\left(\hat{\sigma}^2_{\delta,B-WLS}\right)^{-1} \tag{3.28}$$

where

$\boldsymbol{W}$ is a $(k \times n)$ matrix of weights

This matrix is used to compute the least squares estimate of the regional skew model parameters, $\widehat{\boldsymbol{\beta}}_{WLS}$:

$$\widehat{\boldsymbol{\beta}}_{WLS} = \boldsymbol{W}\widehat{\boldsymbol{Y}} \tag{3.29}$$

***Step Three: Bayesian GLS analysis of model precision***

Given the model parameters and weights calculated in step two, a Bayesian GLS (B-GLS) analysis is conducted to determine the precision of the regional skew model parameters. A procedure similar to that discussed in step 2 is conducted to find the GLS estimate of the model error variance, $\hat{\sigma}^2_{\delta,B-GLS}$. For this analysis, equation (3.24) has been altered to the form

$$f\left(\sigma^2_{\delta,B-GLS}\middle|\widehat{\boldsymbol{Y}},\widehat{\boldsymbol{\beta}}_{WLS}\right) \propto \left|\boldsymbol{\Lambda}_{GLS}\left(\sigma^2_{\delta,B-GLS}\right)\right|^{-0.5} \cdot \exp\left[-0.5\left(\widehat{\boldsymbol{y}} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{WLS}\right)^T\boldsymbol{\Lambda}_{\text{GLS}}\left(\sigma^2_{\delta,B-GLS}\right)^{-1}\left(\widehat{\boldsymbol{y}} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{WLS}\right)\right]\xi\left(\sigma^2_{\delta,B-GLS}\right) \tag{3.30}$$

Here $\boldsymbol{\Lambda}_{GLS}\left(\hat{\sigma}^2_{\delta,GLS}\right)$ is the GLS covariance matrix, defined as

$$\boldsymbol{\Lambda}_{GLS}\left(\hat{\sigma}^2_{\delta,B-GLS}\right) = \hat{\sigma}^2_{\delta,B-GLS}\boldsymbol{I} + \Sigma(\widehat{\boldsymbol{Y}}) \tag{3.31}$$

where

$\sum(\widehat{\boldsymbol{Y}})$ is an $(n \times n)$ matrix, with the sampling variances of the unbiased at-site skew coefficients, $Var[\widehat{\boldsymbol{y}}_i]$, on the diagonal and sampling covarinaces of the unbiased at-site skew coefficients on the off-diagonal.

The procedure for estimating sampling covariance is discussed in more detail in section 3.2.3. Given the B-GLS estimate of $\sigma^2_{\delta,GLS}$, $\boldsymbol{\Lambda}_{GLS}\left(\hat{\sigma}^2_{\delta,B-GLS}\right)$ is calculated using equation (3.31). The GLS covariance matrix for $\widehat{\boldsymbol{\beta}}_{WLS}$, $\sum(\widehat{\boldsymbol{\beta}}_{WLS})$, is given by:

$$\boldsymbol{\Sigma}\left(\widehat{\boldsymbol{\beta}}_{WLS}\right) = \boldsymbol{W}\boldsymbol{\Lambda}_{GLS}\left(\hat{\sigma}^2_{\delta,B-GLS}\right)\boldsymbol{W}^T \tag{3.32}$$

***Variance of Prediction***

As defined in section 3.2.1, the variance of prediction describes the precision of the regional model. Since the hybrid WLS/GLS procedure uses WLS to generate the model and GLS to estimate its precision, the WLS/GLS variance of prediction equations use the WLS weights, the GLS estimate of the model error variance, and the GLS covariance matrix [Lamontagne et al., 2012, Appendix 3]. Thus:

$$VP_{new}(i) = E\left[\sigma^2_{\delta,B-GLS}\right] + \boldsymbol{x}_i\boldsymbol{W}\boldsymbol{\Lambda}_{GLS}\left(\hat{\sigma}^2_{\delta,B-GLS}\right)\boldsymbol{W}^T\boldsymbol{x}_i^T \tag{3.33}$$

$$AVP_{new} = \frac{1}{n}\sum_{i=1}^{n} VP_{new}(i) \tag{3.34}$$

and,

$$VP_{old}(i) = E\left[\sigma^2_{\delta,B-GLS}\right] + \boldsymbol{x}_i\boldsymbol{W}\boldsymbol{\Lambda}_{GLS}\left(\hat{\sigma}^2_{\delta,B-\,GLS}\right)\boldsymbol{W}^T\boldsymbol{x}_i^T - 2\hat{\sigma}^2_{\delta,B-GLS}\boldsymbol{x}_i\boldsymbol{W}\boldsymbol{e}_i \tag{3.35}$$

$$AVP_{old}(i) = \frac{1}{n}\sum_{i=1}^{n} VP_{old}(i) \tag{3.36}$$

where,

$e_i$ is an $(n \times 1)$ vector with one at the $i^{\text{th}}$ row and zero otherwise

***Leverage and Influence***

Since a WLS framework was used in this study to select the regional skew model coefficients, the WLS covariance matrix, $\boldsymbol{\Lambda_{WLS}}$, should be used to determine the leverage for each site. Following the same framework as Tasker and Stedinger [1989], the correct leverage values for this study are provided by

$$\boldsymbol{H^*_{WLS}} = \boldsymbol{X}\left\{\boldsymbol{X^T\Lambda^{-1}_{WLS}X}\right\}^{-1}\boldsymbol{X^T\Lambda^{-1}_{WLS}}$$

or

$$\boldsymbol{H^*_{WLS}} = \boldsymbol{XW} \tag{3.37}$$

The second expression, $\boldsymbol{XW}$, used the definition of the WLS weights, $\boldsymbol{W}$, provided in equation (3.28). The leverage for site $i$ is provided by the $i^{\text{th}}$ diagonal element of $\boldsymbol{H^*_{WLS}}$.

Noting that $\frac{h_{ii}}{1-h_{ii}}$ equals $var(\hat{y}_i)/var(\hat{\epsilon}_i)$, Tasker and Stedinger [1989]'s generalized Cook's D can be decomposed into two parts: the squared residual divided by its variance and the ratio $var(\hat{y}_i)/var(\hat{\epsilon}_i)$ (see equation (3.15) for the OLS case of this decomposition). Given this decomposition, Weisberg [1985; pg 120] explains an observation's influence comes from two sources: the lack of fit of the model (squared standardized residual) and the observation's leverage or potential [$var(\hat{y}_i)/var(\hat{\epsilon}_i)$]. Because the model parameter estimates were determined through a WLS analysis, the observation's potential should be represented in terms of a WLS analysis. As the precision of the model was described with a GLS analysis, the lack of fit of the model at an observation should be expressed in terms of a GLS analysis. Thus, the correct measure of influence for this study is provided by [Veilleux, 2012; Veilleux et al., 2011; Lamontagne, 2012 Appendix 3]:

$$D_i^{WG} = \left(\frac{1}{k}\right)\left(\frac{var[\hat{\gamma}_i|WLS\ Model]}{var[\hat{\varepsilon}_i|WLS\ Model]}\right)\left(\frac{\varepsilon_i^2}{var[\hat{\varepsilon}_i|GLS\ Model]}\right)$$

or

$$D_i^{WG} = \left(\frac{1}{k}\right)\left(\frac{h^*_{WLS,ii}}{1-h^*_{WLS,ii}}\right)\left(\frac{\varepsilon_i^2}{var[\hat{\varepsilon}_i|GLS\ Model]}\right) \tag{3. 38}$$

where

$h^*_{WLS,ii}$ is the $i^{\text{th}}$ diagonal element of $\boldsymbol{H}^*_{WLS} = \boldsymbol{H}_{WLS} * \boldsymbol{\Lambda}_{WLS}$

and

$$\begin{aligned} var[\hat{\gamma}_i|WLS\ Model] &= (\boldsymbol{H}_{WLS})\boldsymbol{\Lambda}_{WLS}(\boldsymbol{H}_{WLS})^T \\ &= (\boldsymbol{X}\boldsymbol{W}_{WLS})\boldsymbol{\Lambda}_{WLS}(\boldsymbol{W}^T_{WLS}\boldsymbol{X}^T) \\ &= \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Lambda}^{-1}_{WLS}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Lambda}^{-1}_{WLS}\boldsymbol{\Lambda}_{WLS}\boldsymbol{\Lambda}^{-1}_{WLS}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Lambda}^{-1}_{WLS}\boldsymbol{X})^{-1}\boldsymbol{X}^T \\ &= \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Lambda}^{-1}_{WLS}\boldsymbol{X})^{-1}\boldsymbol{X}^T \\ &= \boldsymbol{X}\boldsymbol{W}\boldsymbol{L} = (\boldsymbol{H}^*_{WLS})\boldsymbol{\Lambda}_{WLS} \end{aligned} \tag{3.39}$$

$$\begin{aligned} var[\hat{\varepsilon}|WLS\ Model] &= E[(\boldsymbol{\gamma} - (\boldsymbol{H}^*_{WLS})\boldsymbol{\gamma})(\boldsymbol{\gamma} - (\boldsymbol{H}^*_{WLS})\boldsymbol{\gamma})^T] \\ &= \boldsymbol{\Lambda}_{WLS} - (\boldsymbol{H}^*_{WLS})\boldsymbol{\Lambda}_{WLS} - \boldsymbol{\Lambda}_{WLS}(\boldsymbol{H}^*_{WLS})^T \\ &\quad + (\boldsymbol{H}^*_{WLS})\boldsymbol{\Lambda}_{WLS}(\boldsymbol{H}^*_{WLS})^T \end{aligned} \tag{3. 40}$$

$$\begin{aligned} var[\hat{\varepsilon}|GLS\ Model] &= E[(\boldsymbol{\gamma} - (\boldsymbol{H}^*_{WLS})\boldsymbol{\gamma})(\boldsymbol{\gamma} - (\boldsymbol{H}^*_{WLS})\boldsymbol{\gamma})^T] \\ &= \boldsymbol{\Lambda}_{GLS} - (\boldsymbol{H}^*_{WLS})\boldsymbol{\Lambda}_{GLS} - \boldsymbol{\Lambda}_{GLS}(\boldsymbol{H}^*_{WLS})^T \\ &\quad + (\boldsymbol{H}^*_{WLS})\boldsymbol{\Lambda}_{GLS}(\boldsymbol{H}^*_{WLS})^T \end{aligned} \tag{3. 41}$$

For brevity, $\boldsymbol{\Lambda}_{WLS} = \boldsymbol{\Lambda}_{WLS}\left(\hat{\sigma}^2_{\delta,B-WLS}\right)$, and $\boldsymbol{\Lambda}_{GLS} = \boldsymbol{\Lambda}_{GLS}\left(\hat{\sigma}^2_{\delta,B-GLS}\right)$.

The application of this new influence measure is discussed in Chapter 4.

***Section 3.2.3 Cross-Correlation Models***

Martins and Stedinger [2002] developed a relation between the cross-correlation of the at-site sample skew coefficient estimators ($\hat{\rho}\left(\hat{\gamma}_i, \hat{\gamma}_j\right)$) and the cross-correlation of concurrent annual peaks between two sites ($\hat{\rho}_{ij}$) through Monte Carlo experimentation. This relationship was used in this study, and is provided by

$$\hat{\rho}(\hat{\gamma}_i, \hat{\gamma}_j) = sign(\rho_{ij}) cf_{ij} |\hat{\rho}_{ij}|^{\kappa} \quad (3.42)$$

where

$\hat{\rho}_{ij}$ is the cross-correlation of concurrent annual peaks at sites $i$ and $j$
$\kappa$ is a constant between 2.8 and 3.3
$cf_{ij}$ is a factor which accounts for the difference in sample size between basins $i$ and $j$.

$cf_{ij}$ is defined as

$$cf_{ij} = N_{ij} / \sqrt{(N_{ij} + N_i)(N_{ij} + N_j)} \quad (3.43)$$

where

$N_{ij}$ is the length of the concurrent record
$N_i$, $N_j$ is the length of record for basins $i$ and $j$ respectively

Given $\hat{\rho}(\hat{\gamma}_i, \hat{\gamma}_j)$ and an estimate of the sampling variance of $\hat{\gamma}$ for each site, $\sigma^2_{\hat{\gamma}}$, the covariance matrix can be calculated. Recall that covariance can be expressed:

$$cov(\hat{\gamma}_i, \hat{\gamma}_j) = \hat{\rho}(\hat{\gamma}_i, \hat{\gamma}_j) \sigma_{\hat{\gamma}_i} \sigma_{\hat{\gamma}_j} \quad (3.44)$$

where

$\sigma_{\hat{\gamma}_i}$, $\sigma_{\hat{\gamma}_j}$ are the standard deviations of the at-site sample skew coefficient of for basins $i$ and $j$ respectively

Use of the sample cross-correlation of the concurrent annual maximum flows in equation (3.43) is not recommended. Tasker and Stedinger [1989] state that "sample estimates of $\hat{\rho}_{ij}$ are imprecise given short record lengths usually encountered," which "can result in sets of $\hat{\rho}_{ij}$ which make neither hydrologic nor statistical sense." Instead they recommend use of a smooth cross-correlation model based on the distance between basin gauging stations. Their model has the form:

$$\hat{\rho}_{ij} = exp\left\{\left[\frac{d_{ij}}{\alpha d_{ij} + 1}\right] \ln\ \theta\right\} \quad (3.45)$$

where

$d_{ij}$ is the distance between basin gauging stations $i$ and $j$
$\alpha$ and $\theta$ are model parameters such that $\alpha > 0$ and $0 < \theta < 1$

Reis et al. [2005] considered both a constant and a linear distance cross-correlation model in separate case studies. The linear distance model had the form:

$$\hat{\rho}_{ij} = 1 - 0.00291 d_{ij} \tag{3.46}$$

where

$d_{ij}$ is the distance between basin gauging stations $i$ and $j$ in kilometers such that $d_{ij} < 300\ km$

The assumption of normal additive errors is a key assumption of many linear regression diagnostic statistics and is common in least squares regression [Weisberg, 1985, Draper and Smith, 1967]. A normal distributed variate can take values on the interval [-∞, +∞], while cross-correlation is limited to the interval [-1, +1]. Thus, Gruber and Stedinger [2008] employ a Fisher Z transformation to map the sample cross-correlations to the real numbers on the interval [-∞, +∞]. The Fisher Z transformation is provided by [Kendall and Stewart, 1961]:

$$\text{Fisher Z} = \frac{1}{2} \ln \left[ \frac{1+r}{1-r} \right] \tag{3.47}$$

where

$r$ is the sample cross-correlation between two basins

Instead of modeling sample cross-correlation, Gruber and Stedinger [2008] recommend modeling Fisher Z.

Tasker and Stedinger [1989] advise that equation (3.45) could be improved with different functional forms or through the addition of different basin characteristics. Gruber and Stedinger [2008] experiment with many functional forms of both distance and drainage area for the Southeast study, using ordinary least squares regression. A four parameter exponential decay function of distance between basin centroids was shown to perform best:

$$Z_{ij} = \beta_0 + \exp\left(\beta_1 + \beta_2\left(\frac{d_{ij}^{\beta_3} - 1}{\beta_3}\right)\right) + \varepsilon \tag{3.48}$$

where

$\beta_0$, $\beta_1$, and $\beta_2$are model parameters determined by OLS regression
$d_{ij}$ is the distance between the centroids of basin $i$ and $j$
$\varepsilon$ is a random error

An important advance in equation (3.48) is the use of basin centroids rather than gauging stations to determine basin-to-basin distance. The basin gauging station is almost always located at the basin outlet (an exception in this study was the use of upstream and downstream gauges to estimate a river's flood record at some point). Use of basin gauging station location by Tasker and Stedinger [1989] and Reis et al. [2005] might be problematic as adjacent gauging stations might drain basins composed of areas spanning a great distance. Cottonwood Creek near Cottonwood (study site 3) and Battle Creek near the Coleman fish hatchery (study site 5) are an example of such a pairing (see Figure 1.1). While their gauging stations are adjacent (less than 5 miles apart), they drain opposite sides of the Central Valley. The distance between their centroids better reflects this (51 miles apart).

In this study a variety of functional forms for relating the Fisher Z transformed cross-correlation to the distance between basin centroids were experimented with. Ultimately a three parameter exponential decay model having the following form appeared to best represent the cross-correlations in this study:

$$Z_{ij} = a + \exp(b - c \times d_{ij}) \tag{3.49}$$

Further discussion of correlation model selection for this study can be found in Chapter 4.

### *Section 3.3: Redundant Basins*

#### *Section 3.3.1: Redundant Basins and Redundancy Metrics*

When conducting a regional hydrologic regression study, it is good to ensure that no two basins are redundant. Redundancy occurs when one basin is nested within another so that they represent largely the same area and are not two spatially independent observations. In this case, it is likely that they will experience very similar hydrologic events so that their model error will be correlated. Thus, retention of both basins in a regional skew coefficient model is statistically incorrect [Gruber and Stedinger, 2008]. This study involved only 50 sites, many of which were familiar to researchers, so that redundancy was not of great concern. Other studies might involve sufficiently many sites so that such manual examination of basins to ensure autonomy is not feasible. For example, the California instantaneous annual peak study included 146 basins across California [Parrett et al., 2011]. In such cases, metrics can be used to identify basins which might be redundant so that the researcher can examine those few cases more closely. The California and Southeast instantaneous annual peak skew studies utilized two metrics: normalized distance and drainage area ratio. Normalized distance, $ND$, is defined as [Gruber and Stedinger, 2008]:

$$ND = \frac{d_{ij}}{\sqrt[4]{DA_i * DA_j}} \tag{3.50}$$

where

$d_{ij}$ is the distance between the centroids of basins $i$ and $j$
$DA_i$, $DA_j$ are the drainage areas of basins $i$ and $j$ respectively

Drainage area ratio, $DAR$, is defined as [Gruber and Stedinger, 2008]:

$$DAR = \max\left(\frac{DA_i}{DA_j}, \frac{DA_j}{DA_i}\right) \tag{3.51}$$

To understand these metrics more intuitively consider a delineated watershed as roughly triangular. Admittedly this is an over-simplification, but it holds true if one considers that watersheds typically expand in breadth as one travels up the main channel from the outlet. Given this approximation, it is clear that the basin centroid will typically lay well within the basin, likely farther from the narrow outlet than the broad farthest reach. Given roughly triangular watersheds, the distance between two neighboring basins' centroids should increase as a function of their drainage areas. For smaller basins, one would expect much less distance between neighboring basins' centroids than for larger basins. To reflect this, $ND$ is the centroid-to-centroid distance scaled by the product of the drainage areas.

Gruber and Stedinger [2008] point out that nested basins are not necessarily redundant. In the case of a very small basin nested within a much larger basin, the difference in their size and characteristics might be enough to ensure only weak cross-correlation of the skew coefficient. $DAR$ is a measure of how similar two basins are in size, the idea being to determine the extent to which two basins might represent the same geographic area and are, for the purposes of a regional skew study, redundant. Gruber and Stedinger [2008] advises that $DAR$ less than or equal to 5 and $ND$ less than or equal to 0.5 be used as a guideline for determining basins which might be nested and redundant.

If the size of all basins in a regional hydrologic study are on the same order of magnitude, the $ND$ and $DAR$ metrics work well. However, problems arise if some basins are many times larger than others. In this case, several basins might be nested within a single large basin without violating the $ND$ and $DAR$ provided above. Parrett

et al. [2011] document such a case, in which several basins were nested within the Sacramento River at Keswick and the Feather River at Oroville Dam basins, despite $ND$ and $DAR$ being greater than 0.5 and 5 respectively. In that case both large basins were dropped from the skew study so that the smaller basins could be used.

For such cases, a new redundancy metric was developed and applied to this study. Standardized distance, $SD$, is defined as

$$SD = \frac{d_{ij}}{\sqrt{(DA_i + DA_j)/2}} = \frac{\sqrt{2}d_{ij}}{\sqrt{DA_i + DA_j}} \quad (3.\ 52)$$

By using the sum rather than the product of the drainage areas, the $SD$ statistic is more sensitive to nested basins of vastly different sizes than $ND$. The $\sqrt{2}$ factor allows $SD$ and $ND$ to have roughly the same scale.

***Section 3.3.2: Model Error Correlation***

WLS and GLS regression, as presented by Stedinger and Tasker [1985], divides the regression error into two elements: the error due to the use of an imperfect model and the sampling error due to short record length. In a WLS analysis, both model error and sampling error are assumed to be uncorrelated between study basins, so the off-diagonal elements of the covariance matrix $\mathbf{\Lambda}$ are zero. In the Stedinger and Tasker [1985] GLS framework, error correlation is attributed solely to sampling error. This means that the off-diagonal elements of $\mathbf{\Lambda}$ are estimates of the covariance of the at-site sample skew coefficients. Model errors are still assumed to be independent, with mean zero. This assumption is not universally accepted [Kjeldsen and Jones, 2009].

The Flood Estimation Handbook (FEH), which provides guidelines for flood frequency in the United Kingdom, involves the use of an index flood when conducting a flood frequency analysis at an ungauged site [Institute of Hydrology, 1999]. To generate an estimate of the index flood for ungauged sites, Kjeldsen and Jones [2009] utilize a variation of GLS regression. Their formulation of GLS allows for the possibility of model error correlation, indicating the regression model systematically fails to predict the index floods at basins in a predictable way which can be modeled. They attribute this systematic failure to the inability of a simple regression model based on basin characteristics to fully reflect the complexity of basin dynamics [Kjeldsen and Jones, 2009; Kjeldsen and Jones, 2006].

Kjeldsen and Jones [2007] present empirical evidence of model error correlation in an index flood study for the United Kingdom. Systematic failure of a regression model might be caused in various ways besides the complexity of the natural environment. Veilleux [2009] offers a discussion of model error correlation in the context of redundant basins. Her argument is that one would expect model error correlation if two basins are redundant as they represent essential the same watershed. Kjeldsen and Jones [2009] did not remove nested basins and acknowledge that nested basins are an "intuitive" explanation of model error correlation. Veilleux [2009] argues that removing redundancies improves the regression analysis with only marginal loss of data, since redundant basins represent nearly the same watersheds.

Another cause of model error correlation is the use of an insufficient model. Systematic failure of a model can often indicate that some important explanatory variable has been neglected. The assertion by Kjeldsen and Jones [2009] that a simple

regression model cannot fully capture the true complexity of the natural environment is not disputed by the Stedinger and Tasker [1985] GLS framework; rather the model errors discussed by Stedinger and Tasker [1985], Reis et al. [2005], and this thesis are caused by precisely this. The departure is in the assertion that model errors are correlated and that this correlation might be described in terms of distance between basins [Kjeldsen and Jones, 2006, 2007]. The rational for modeling cross-correlation of model errors as a function of distance between basins is that basins which are close exhibit similar hydrologic characteristics which impact the flood statistic of interest and have been neglected from the regression. While the neglected explanatory variable might be impossible to determine precisely, it seems likely that a surrogate variable could be developed. As Kjeldsen and Jones [2009] state that they have rigorously examined all reasonable combinations of explanatory variables and are still experiencing model error correlation, nested redundant basins are more likely the cause of this correlation, or else use of an inadequate correlation model for sampling error of the index flood.

***Conclusion:***

This chapter provides the theoretical framework for regional hydrologic regression based on the Bayesian GLS procedure proposed by Reis et al. [2005]. A variety of diagnostic statistics for the Bayesian GLS analysis are reported and explained. Additionally, a new WLS/GLS procedure is introduced. This procedure uses Bayesian WLS to estimate the model parameters, but Bayesian GLS to assess its precision. This methodology was developed because the high correlation between skew sampling errors in California resulted in complex GLS regression weights which

were not justifiable [Parrett et al., 2011; Lamontagne et al., 2012]. Finally, this chapter explores the potential impact of redundant basins on hydrologic regression and presents a new metric, standardized distance ($SD$), to detect them.

# REFERENCES


Brutsaert, W. (2005), Hydrology: An Introduction, Cambridge University Press, New York, NY., pp. 509-550.

Chow, V. T., D. R. Maidment, and L. W. Mays. (1988). *Applied Hydrology*, McGraw-Hill, New York.

Draper, N.R. and Smith, H. (1967). *Applied Regression Analysis*. John Wiley & Sons, Inc., New York, N.Y.

Eash, D.A., Barnes, K.K., and Veilleux, A.G., 2013, Methods for estimating annual exceedance-probability discharges for streams in Iowa, based on data through water year 20120: U.S. Geological Survey Scientific Investigations Report 2013-5086, 63 p. with appendix.

Feaster, T.D., Gotvald, A.J., and Weaver, J.C., 2009, Magnitude and frequency of rural floods in the southeastern United States, 2006—Volume 3, South Carolina: U.S. Geological Survey Scientific Investigations Report 2009-5156, 226 p.

Griffis, V. W., and J. R. Stedinger, (2007), The Use of GLS Regression in Regional Hydrologic Analyses, J. of Hydrology, 344(1-2), 82-95, [doi:10.1016/j.jhydrol.2007.06.023].

Griffis, V.W., and J. R. Stedinger, (2009), The Log-Pearson Type 3 Distribution and its Application in Flood Frequency Analysis, 3. Sample Skew and Weighted Skew Estimators, *J. of Hydrol. Engineering* 14(2), pp. 121-130.

Gruber, A.M., D.S. Reis Jr., and J. R. Stedinger (2007), Models of regional skew based on Bayesian GLS regression, *World Environmental & Water Resources Conference-Restoring out Natural Habitat*, edited by K.C. Kabbes, Tampa, Florida
May 15-18, Paper 40927-3285

Gruber, A.M. and J.R. Stedinger, (2008), Models of LP3 Regional Skew, Data Selection and Bayesian GLS Regression, Paper 596, World Environmental and Water Resources Congress – Ahupua'a, Babcock, R.W. and R. Watson editors, Honolulu, Hawai'i, May 12-16.

Gotvald, A.J., Feaster, T.D., and Weaver, J.C., 2009, Magnitude and frequency of rural floods in the southeastern United States, 2006—Volume 1, Georgia: U.S. Geological Survey Scientific Investigations Report 2009–5043, 120 p.

Institute of Hydrology (1999), Flood Estimation Handbook, Wallingford, U.K.

Hardison, C.H. (1974)."Generalized Skew Coefficents of Annual Floods in the United States and Their Application." *Water Resources Research* 10, no. 5: 745-752.

Interagency Advisory Committee on Water Data, 1982, Guidelines for determining flood-flow frequency, Bulletin #17B of the Hydrology Subcommittee, Office of Water Data Coordination: U.S. Geological Survey, Reston Virginia, 183 p. Available at *http://water.usgs.gov/osw/bulletin17b/dl_flow.pdf*

Kendall, M.G. and A. Stuart (1961), *The Advanced Theory of Statistics*, Vol 2, Hafner Publishing Company, New York.

Kjeldsen, T. R., and D. A. Jones (2006), Prediction uncertainty in a median-based index flood method using L moments, Water Resour. Res., 42, W07414, doi:10.1029/2005WR004069.

Kjeldsen, T. R., and D. A. Jones (2007), Estimation of an index flood in the UK, Hydrol. Sci. J., 52, 86– 98, doi:10.1623/hysj.52.1.86.

Kjeldsen, T.R., and D.A. Jones (2009), An exploratory analysis of error components in hydrological regression modeling, *Water Resour. Res.*, 45, W02407, doi:10.1029/2007WR006283.

Kroll, C.N., Stedinger, J.R., 1998. Regional hydrologic analysis: ordinary and generalized least squares revisited. Water Resour. Res. 34 (1), 121–128.

Kroll, C.N., Stedinger, J.R., 1999. Development of regional regression relationships with censored data. Water Resour. Res. 35 (3), 775–784.

Lamontagne, J.R., Stedinger, J.R., Berenbrock, Charles, Veilleux, A.G., Ferris, J.C., and Knifong, D.L., 2012, Development of regional skews for selected flood durations for the Central Valley Region, California, based on data through water year 2008: U.S. Geological Survey Scientific Investigations Report 2012–5130, 60 p.

Mason, R., 2013."A Partial Summary of 2012 USGS Activities of Interest to the FHWA and State Highway Agencies." 91st TRB Annual Meeting, AFB60 Committee Meeting. Available at http://water.usgs.gov/osw/pubs/2012_FHWA_USGS.pdf

Martins, E.S., and Stedinger, J. R., 2002, Cross-correlation among estimators of shape: Water Resources Research, v. 38, no. 11, 1252, doi: 10.1029/2002WR001589

Parrett, C., A. Vellieux, , J. R. Stedinger, N. A. Barth, D. Knifong, , and J.C. Ferris, 2011. Regional Skew for California and Flood Frequency for Selected Sites in the Sacramento-San Joaquin River Basin Based on Data through Water Year 2006, OFR, U.S. Geological Survey.

Reis, D.S., Jr., Stedinger, J.R., and Martins, E.S., 2005, Bayesian generalized least squares regression with application to the log Pearson type III regional skew estimation: Water Resources Research, 41, W10419, doi:10.1029/2004WR003445.

Stedinger, J. R., 1983, Estimating a Regional Flood Frequency Distribution: Water Resources Research, v. 19, no. 2, p. 503-510.

Stedinger, J. R., and Tasker, G. D., 1985, Regional hydrologic analysis, 1, ordinary, weighted and generalized least squares compared: Water Resources Research ,v. 21, no. 9, p. 1421-1432. [with correction, Water Resources Research, v. 22, no. 5, p. 844, 1986.]

Stedinger, J.R., and Tasker, G.D., 1986a, Correction to Regional hydrologic analysis 1, ordinary, weighted and generalized least squares compared. Water Resources Research. 22 (5), 844.

Stedinger, J.R., and Tasker, G.D., 1986b, Regional hydrologic analysis 2. Model-error estimators, estimation of sigma and log-Pearson type 3 distributions. Water Resources Research. 22 (10), 1487-1499.

Tasker, G .D., and Stedinger, J. R., 1986, Regional skew with weighted LS regression: Journal of Water Resources Planning and Management, ASCE, v.112, no. 2, p. 225–237.

Tasker, G.D., and J.R. Stedinger, 1989. An Operational GLS Model for Hydrologic Regression, Journal of Hydrology, 111(1-4), 361–375.

Weaver, J.C., Feaster, T.D., and Gotvald, A.J., (2009), Magnitude and frequency of rural floods in the southeastern United States, 2006—Volume 2, North Carolina: U.S. Geological Survey Scientific Investigations Report 2009–5158, 113 p.

Weisberg, S. (1985). *Applied Linear Regression*. John Wiley & Sons, Inc., New York, N.Y.

Veilleux, A. G. 2009. "Bayesian GLS Regression for Regionalization of Hydrologic Statistics, Floods and Bulletin 17 Skew." MS thesis, School of Civil and Environmental Engineering, Cornell Univ., Ithaca, N.Y.

Veilleux, A. G., J. R. Stedinger, J. R. Lamontagne, (2011), Bayesian WLS/GLS Regression for Regional Skewness Analysis for Regions with Large Cross-Correlations among Flood Flows, *World Environmental and Water Resources Conference-Bearing Knowledge for Sustainability*, edited by R. E. Beighley II and M. W. Killgore, Palm Springs, California.

Veilleux, A.G. "Re: discordancy impacts and request for digitial file summary of revised map skew." E-mail message to Richard Vogel, August 6, 2013.

# CHAPTER 4

# REGIONAL ANALYSIS OF CALIFORNIA LOG-SPACE SKEW COEFFICENTS FOR D-DAY MAXIMA

This chapter describes the application of the new regional skew analysis methodology developed in Chapter 3to annual maximum rainfall flood records of five durations in the state of California. The specific basins used in this study, as well as a general discussion of their hydrology can be found in Chapter 1. Section 4.1 discusses the observed log-space skew coefficients of the flood records used in the study. Section 4.2 describes the difficulties encountered when attempting to apply traditional regionalization techniques to this study, and Section 4.3 provides justification for the use of the methodology described in Chapter 3 for this study. Section 4.4 presents the results of the case study. Section 4.5 discusses the application of new leverage and influence statistics developed for this study, and Section 4.6 discusses the development of the non-linear elevation terms utilized in the regional skew study and examines its parameterization.

## *Section 4.1: Observed log-space skew coefficient*

Before a regional skew analysis could be conducted, reasonable estimates of the at-site sample skew coefficient had to be computed for each study site and each duration. The presence of low-outliers in a flood series can make obtaining reliable estimates of the sample skew difficult. Bulletin 17B defines an outlier as an observation "which departs from the trend of the rest of the data [IACWD, 1982]." Various low-outlier identification and censoring procedures have been proposed and

shown to improve flood frequency results [IACWD, 1982; Cohn et al., 2001; Griffis et al., 2004]. These procedures were discussed in more detail in Chapter 2 of this thesis.

This study used an iterative manual identification procedure in which low outliers were visually identified by observing the fitted distribution on a probability plot. When the fitted distribution failed to accurately represent the data, and particularly when it failed to describe the largest observations of a flood record, the smallest observations were incrementally censored and the distribution was refit to the retained data using the Expected Moments Algorithm [Cohn et al., 2001].

For the first iteration of the procedure the Grubbs-Beck lower bound threshold recommended by Bulletin 17B was used [IACWD, 1982]. In cases where additional censoring was deemed necessary, the EMA censoring threshold was adjusted to be less than the smallest retained observation. In no case was an observation identified as a low-outlier by the Grubbs-Beck test retained in the analysis. For more details concerning the censoring procedure and the extent of censoring in this study, please see Section 2.4.

Flood series were analyzed for five different flow durations. In most instances the smallest observations in a flood record corresponded to the same hydrologic event across all durations, i.e. the 1-day duration flood usually occurred during the 30-day duration flood. Thus, an effort was made to ensure censoring consistency across durations. In some instances minor inconsistencies were allowed if the fitted distribution described the largest observations of a sample well without further censoring. Censoring too liberally, within reason, should not greatly impact the estimated skew coefficient given the long record lengths in this study. The extent of

censoring for flood records used in this study is discussed at length in Section 2.4.2, and a summary of censoring for individual sites can be found in Table 2.1. The observed log-space skew coefficient for each basin and duration used in this study are plotted in Figure 4.1 in order of ascending 7-day log-space skew and listed in Table A.1-Table A.7 in Appendix A.

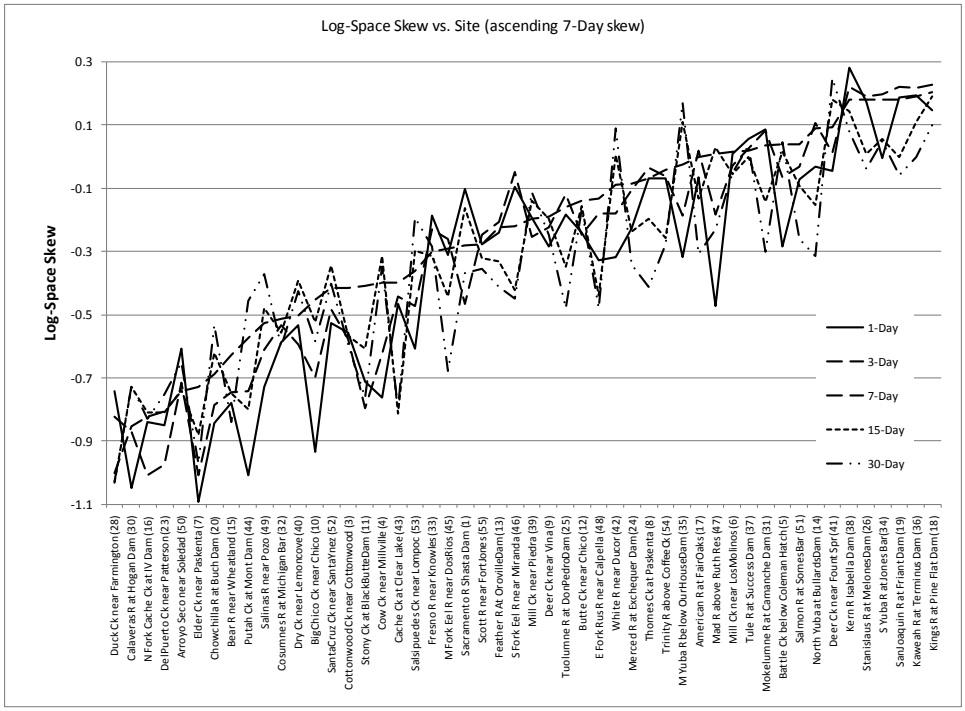


**Figure 4.1:** Log-space skew for rainfall floods all durations vs. site name, in order of ascending 7-day log-space skews.

Note that at-site sample log-space skew coefficients ranged roughly between -1.1 and 0.2. While a fair amount of variation between durations was observed for each basin, skew coefficients for all durations trend together across study basins.

***Section 4.2: Cross-Correlation of annual rainfall floods for California***

Many hydrologic variables, including the at-site sample skew coefficient, are cross-correlated among basins which are geographically close or hydrologically similar [Stedinger and Tasker, 1985]. Regional skew regression methods which do not consider cross-correlation of sample skew coefficients implicitly assume they are independent observations. If the degree of correlation is significant, this assumption might impact the fitted regression parameters and will almost certainly impact the estimated precision of the model [Reis et al., 2005]. This is important in this application, because Bulletin 17B procedures use a weighted average of the regional and sample skew coefficients for flood frequency, in which the weighting is based on the relative precision of the two skew estimates. Thus it is important to estimate the precision of the model well.

No direct estimate of the cross-correlation of the sample skew coefficients of flood records is readily available. Martins and Stedinger [2002] provide a relationship between the cross-correlation of two annual maximum flood series and the cross-correlation of their estimated sample skew coefficients. The cross-correlation of the sample skew coefficients for each duration in this study were estimated from the cross-correlation of the rainfall flood records between each site.

The observed cross-correlation between concurrent rainfall floods of various durations were significantly greater than those observed between concurrent annual instantaneous peak flows in California [Parrett et al., 2011]. This was expected because instantaneous peak flows in California are caused by rain and snowmelt events, while rainfall floods are caused predominantly by rainfall only. This increases

the chances that two annual maxima occur essentially at the same time and thus are likely to be highly correlated. Furthermore, the averaging of flood volumes over longer durations diminished the effect of spatial and temporal variability of hydrologic events, resulting in higher cross-correlations. The cross-correlation between rainfall floods increased with flood duration so that cross-correlation between 30-day floods was nearly always greater than cross-correlation between 1-day floods.

To construct the GLS covariance matrix, Tasker and Stedinger [1989] recommend utilization of a cross-correlation model rather than sample cross-correlations in order to create a covariance matrix which is consistent and hydrologically defensible. Gruber and Stedinger [2008] illustrate that modeling the Fisher Z transformation of the cross-correlation of annual maximum floods is well suited to the skew regression framework and can improve regional skew analysis results. This is discussed in more detail in Section 3.2.

Tasker and Stedinger [1989], Reis et al. [2005], Gruber and Stedinger [2008], and Parrett et al. [2011] model cross-correlation (or Fisher Z transformed cross-correlation) as a function of the distance between basins. This study experimented with various functional transformations of distance to describe Fisher Z transformed cross-correlation between annual maximum rainfall floods for each of the durations. In this study distance was measured between basin centroids.

The precision of the Fisher Z transformed cross-correlation estimator for two sites is a function of the concurrent record [Veilleux, 2009]. By setting a minimum concurrent record length for a site pairing to be included in the correlation model development, only more reliable cross-correlation estimates were utilized. A

concurrent record threshold uses reliable correlation estimates, but also retains enough site pairings for a good analysis. A concurrent record length of at least 50 years was adopted. With a concurrent record threshold of 50 years for the 1-day duration, 624 of 1,225 possible pairings were used, which included 42 of 50 possible sites. Eight sites were excluded from the analysis because they had record lengths less than the 50 years.

Parrett et al. [2011] selected a two parameter exponential decay model for Fisher Z transformed cross-correlations between annual maximum floods in California. The only explanatory variable for their model was distance between basin centroids. The model adopted for this study (Equation (4.1) is a generalization of their model, including an extra, additive constant parameter. It was observed that this model better described the transformed cross-correlation for the rainfall floods considered in this study. Unlike the Parrett et al. [2011] model, this correlation model does not converge to zero at some great distance. It is important to note that this model was intended to provide cross-correlation between study basins which are fairly close, and that sample correlation did not drop to zero in the distance range of interest. The model is:

$$Z_{ij} = a + \exp\left(b - c \times d_{ij}\right) \tag{4.1}$$

where

$Z_{ij}$ is the Fisher Z transformed cross-correlation between concurrent flood records at sites $i$ and $j$.
$a$, $b$, and $c$ are model parameters selected through regression
$d_{ij}$ is the distance between the centroids of basins $i$ and $j$ in miles

An ordinary non-linear least squares regression was performed to find appropriate model parameters for each duration. Non-linear least squares was

necessary because the model form is not linear. A weighted least squares regression would have been appropriate if sampling variances of the Fisher Z statistic varied greatly between pairings. Because the concurrent record lengths were fairly uniform across all pairings, the sampling variance of the Fisher's Z statistic is also fairly uniform across pairings, and thus an ordinary least squares regression should be adequate. Table 4.1 provides the correlation model parameters for the five study durations. All model parameters were significant at the 0.05 significance level. While it is surprising that the model parameters are not ordered in duration, Figure 4.2 shows that model transformed correlation are generally ordered by duration over the range of distances of interest.

**Table 4.1:** Parameters of cross-correlation models for concurrent flood flows for all durations.

| Duration | a | b | c |
|---|---|---|---|
| **1-Day** | 0.38 | 0.15 | 6.05E-03 |
| **3-Day** | 0.38 | 0.22 | 5.62E-03 |
| **7-Day** | 0.38 | 0.26 | 4.96E-03 |
| **15-Day** | 0.36 | 0.31 | 4.58E-03 |
| **30-Day** | 0.41 | 0.28 | 4.80E-03 |

Table 4.2 reports a summary of statistical results from the correlation regression for each duration for both the selected three parameter model and a constant model. The constant model has the form:

$$Z_{ij} = a \tag{4.2}$$

The constant model has an effective record length between 12 and 13 years, depending on the duration. The constant model has a Pseudo $R^2$ of zero for all durations, as it is explaining none of the variability in the Fisher Z-transformed cross-correlations between flood records. The final model has Pseudo $R^2$ values between 68

and 72%, with effective record lengths between 33 and 36 years depending on duration. These effective record lengths are similar to the reported effective record length of the correlation model used in the previous California instantaneous annual maximum study.

**Table 4.2:** Statistical Summary of Regression for the Final Correlation Model (Eqn. 4.1) and the Constant Model (Eqn. 4.2) for all study durations.

| Duration | Model | MEV | Pseudo $R^2$ | ERL |
|---|---|---|---|---|
| 1 | Constant | 0.1011 | 0 | 13 |
| | Final | 0.0320 | 0.68 | 34 |
| 3 | Constant | 0.1103 | 0 | 12 |
| | Final | 0.0330 | 0.70 | 33 |
| 7 | Constant | 0.1125 | 0 | 12 |
| | Final | 0.0311 | 0.72 | 35 |
| 15 | Constant | 0.1156 | 0 | 12 |
| | Final | 0.0299 | 0.74 | 36 |
| 30 | Constant | 0.1146 | 0 | 12 |
| | Final | 0.0309 | 0.73 | 35 |

Figure 4.2 displays the cross-correlation models for each duration included in this study, as well as the cross-correlation models for the California instantaneous peak study [Parrett et al., 2011] and the Southeast instantaneous peak [Veilleux, 2009] study. The Southeast correlation model declines rapidly with increasing distance, while the California instantaneous peak correlation model declines gradually with increasing distance. The d-day rainfall flood correlation models have a similar shape to the California instantaneous peak correlation model, but are larger. As rainfall flood duration increases, so does the cross-correlation for a fixed distance.

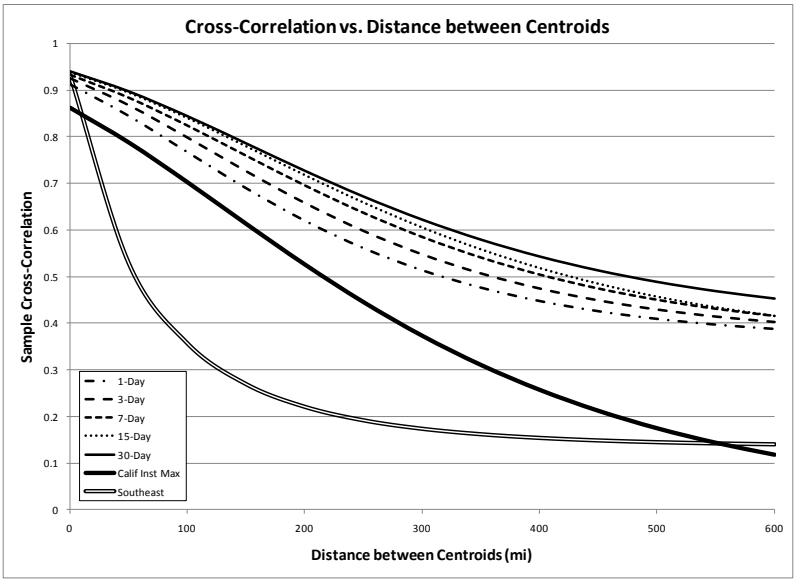


**Figure 4.2:** Models of cross correlation between concurrent annual maximums for all durations as a function of the distance between basin centroids.

***Section 4.3: Methodology adjustment for California rainfall floods***

This section briefly describes the methodical adjustments to the standard Generalized Least Squares (GLS) procedure used to generate regional skew models for rainfall floods in California. For a more theoretical discussion, see Section 3.2: Standard GLS and Hybrid WLS/GLS Procedure.

Figure 4.2 shows that the cross-correlation between annual maximum rainfall floods for all durations were greater than those observed in the previous Southeast study [Veilleux, 2009; Weaver et al. ,2009; Feaster et al., 2009; Gotvald et al., 2009] and the previous instantaneous peak study in California [Parrett et al., 2011]. Parrett et al. [2011] document that high cross-correlation of sampling errors can lead to instability in the regression parameter estimates. This occurs when GLS seeks to exploit the high cross-correlation through a series of complicated weights. Figure 4.3 illustrates regression weight instability by plotting the weights assigned to each site in

this study for a constant model (Equation (4.4)) by OLS, WLS, and GLS versus record length.

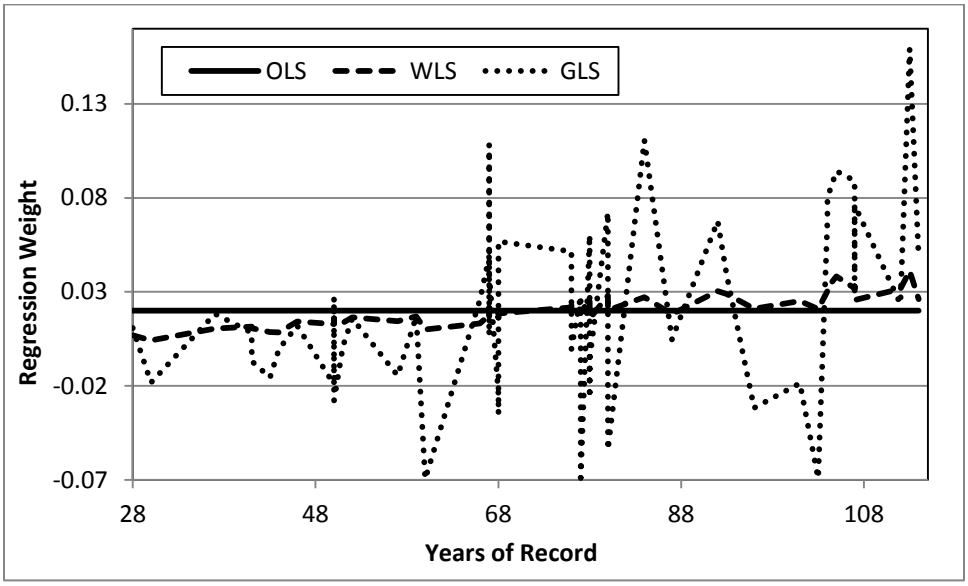


**Figure 4.3:** Comparison of regression weights assigned to study sites (for the constant model) from OLS, WLS, and GLS analyses

The sum of all regression weights must equal one. Note that OLS assigns an equal weight of (1/50) to each site, regardless of record length. WLS, plotted as a dashed line, assigns greater weight to basins with longer record length because the precision of the at-site sample skew coefficient estimator is a function of record length. In this study, WLS assigns a weight of 0.04 for a record length of 114 years versus 0.007 for a record length of 28. The ratio is not quite 2:1. The WLS weights are not perfectly linear in record length because an estimate of the at-site skew is used to find the variance of the skew estimate rather than the true, unknown population skew for that site (see Chapters CHAPTER 2 and CHAPTER 3 for discussion). One would expect the GLS weights to vary somewhat from WLS weights, but the GLS weights should remain reasonable; i.e. not highly negative or positive. While the GLS

weights plotted in Figure 4.3 still sum to one, they depart radically from the WLS weights and are generally erratic.

If the true cross-correlation of the skew coefficients were known, such complex weights might be defensible. Unfortunately the true cross-correlations are not known and are estimated from the cross-correlation of the concurrent rainfall flood records. The precision of these estimates do not justify such complicated weights, so a GLS analysis was judged to be unacceptable [Parrett et al., 2011]. However, a WLS analysis which ignores high cross-correlation would overestimate the precision of the regional model by assuming that sampling errors are independent. Thus, it was necessary to adopt an alternative regression method for the regional skew analysis which utilized a WLS framework to estimate skew model coefficients, but assess the precision of the model using a GLS framework.

***Section 4.3.1: Hybrid OLS/GLS/WLS for California rainfall floods***

Following a procedure similar to that described by Parrett et al. [2011], an OLS/WLS/GLS hybrid procedure was developed for this study of rainfall floods. This process used an OLS regression to create an initial regional skew model. The OLS regional skew coefficients were used to estimate the variance of the at-site sample skew coefficients for each site. With these variances, a WLS regression was used to find regional skew coefficient models for each duration. The precision of these models was then estimated using a Bayesian-GLS procedure. These steps are discussed in greater detail in Section 3.2: Standard GLS and Hybrid WLS/GLS Procedure.

***Section 4.3.2: Sample skew coefficient variance for records experiencing heavy censoring.***

In order to estimate the sampling covariance matrix, it is necessary to obtain reliable estimates of the at-site sample skew coefficient's variance for each basin and duration in the study. For most of the records in this study, these were obtained using the formula provided by Griffis and Stedinger [2009] and repeated in Equation 3.26. This expression was originally obtained by Griffis [2003] through an extensive Monte Carlo analysis involving samples drawn from Pearson Type 3 distributions of record lengths 10 to 150 years. The study involved complete and uncensored samples. No examination of the effects of censoring on the precision of the skew coefficient estimator was made.

If one were to randomly censor observations from a sample, the Griffis and Stedinger [2009] equation would still be valid. On the other hand, censoring of low outliers by removing the smallest observations in a sample is not random and could potentially affect the estimated variance. Thus, an alternative estimate of the sample skew coefficient variance was needed for sites experiencing significant censoring. The expected moments algorithm (EMA) provides an estimate of the variance of the sample skew coefficient. The process used to derive this estimate is documented in the appendix of Cohn et al. [2001].

Forty-eight of the 50 basin records included in this study contained less than five censored observations for all durations, and the Griffis and Stedinger [2009] approximation of the variance of the sample skew coefficient was used. In fact, 46 of the 50 basin records included in this study contained less than four censored observations; the two basins with 4 censored observations are Cache Creek at Clear

Lake (study basin, 43) and North Fork Cache Creek at the Indian Valley dam (study basin, 16). These two basins have record lengths of 83 and 77 years respectively; the censoring for these sites was deemed minor compared to their record length, so the Griffis and Stedinger [2009] formula for the variance of the skew estimator was deemed appropriate.

The two basin records experiencing greater than four censored observations, Santa Cruz Creek at Santa Ynez (study basin, 52) and Putah Creek at the Monticello dam (study basin, 44) warranted more extensive censoring. Depending on duration, four to six observations were censored at Santa Cruz Creek from a record of 78 years, while eleven to twelve observations were censored at Putah Creek from a record of 67 years. This censoring level was deemed too extreme to use the Griffis and Stedinger [2009] formulation, so the EMA estimate of the variance of skew was utilized. Because censoring for most basin records involved one or two observations from record lengths in excess of 70 years, and the heavy censoring sites represented only two basin of 50, the effects of this inconsistency in methodology was thought to be minor. This was checked with a sensitivity analysis. The WLS/GLS regression analysis was rerun without the four heavy censoring sites. There was very little difference in the results, including the average variance of prediction. Thus the handling of these four sites in the OLS/WLS/GLS analysis had very little effect on the regional model or its estimated precision. The change was actually surprising small, but can be understood in that there were 50 sites in all, the Putah Creek and Santa Cruz Creek skews were assigned large sampling variances by EMA and thus small

weights, and that Putah Creek and the two Cache Creek basins are physically very close so that they are highly correlated and do not represent three independent basins.

### *Section 4.4: Regional skew coefficient modeling results*

In this study, five regional skew coefficient models were developed: one for each rainfall flood duration considered. It was expected that flood durations would exhibit different distribution characteristics as floods of different durations can have very different characteristics. On the other hand, the regional skew coefficient models should not be very different, otherwise inconsistencies between flood frequency results for various durations might occur.

This study included 50 basins in and around the Central Valley of California. Record lengths ranged from 28 to 116 years with a mean of 74 years and a standard deviation of 24 years (see Table 1.1 for a list of basins included in the study and their period of record). More than two thirds of basins have record lengths in excess of 65 years. A total of twenty possible explanatory variables were provided by USGS for most basins. These characteristics are described and defined in detail in Chapter 1. The full suite of basin characteristics were not available for all basins, so initial study results were obtained using only 47 basins that had the full set of explanatory variables, while the final models were developed using all 50.

Synthetic basin characteristics, which are combinations or functions of one or more of the provided basin characteristics were also tested. These include several non-linear functions of mean basin elevation. While a variety of basin characteristics were tested, only functions of elevation proved statistically significant. This was also observed by the previous California instantaneous peak study. The importance of

elevation was attributed to the switch from rainfall-only hydrology at low elevations to rain and snowmelt hydrology at high elevations [Parrett et al., 2011].

Basin drainage area is a commonly used characteristic in hydrologic models. Figure 4.4 plots sample skew coefficients for each study site in order of ascending drainage area. Note that no clear trend is exhibited in the plot, indicating that basin drainage area is likely a poor explanatory variable for regional skew.

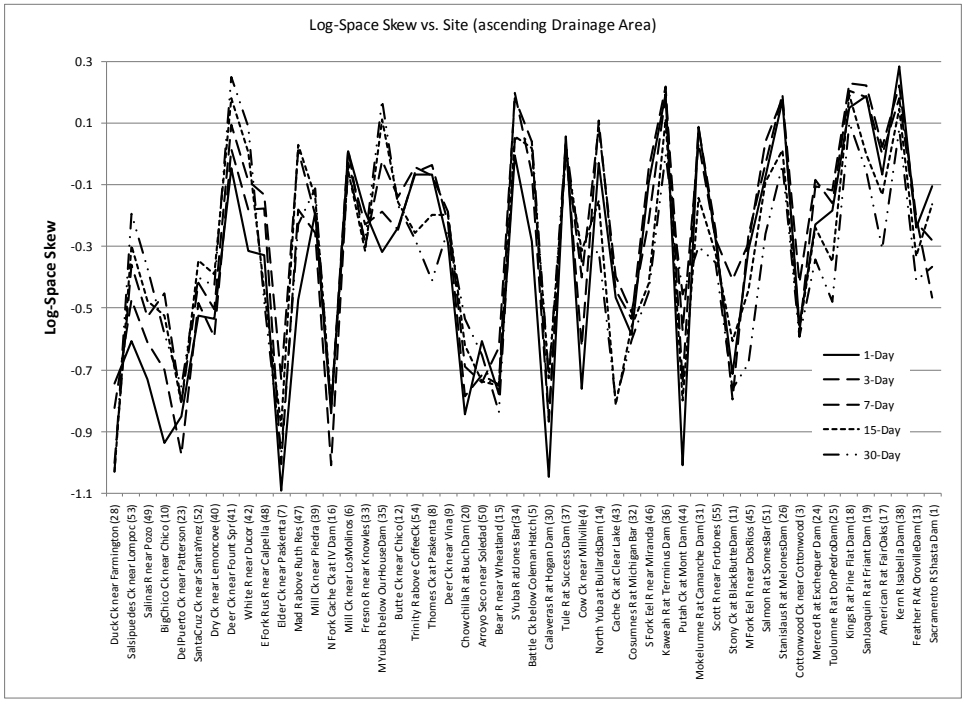


**Figure 4.4:** Observed at-site sample skew coefficients versus site, ascending basin drainage area.

Figure 4.5 displays the skew coefficient for each study basin and duration, sorted by mean basin elevation. Note a clear non-linear trend with elevation, in which low elevation skew coefficients vary about one mean, while high elevation skew coefficients vary about a less negative mean, with a brief and rapid transition zone in between.

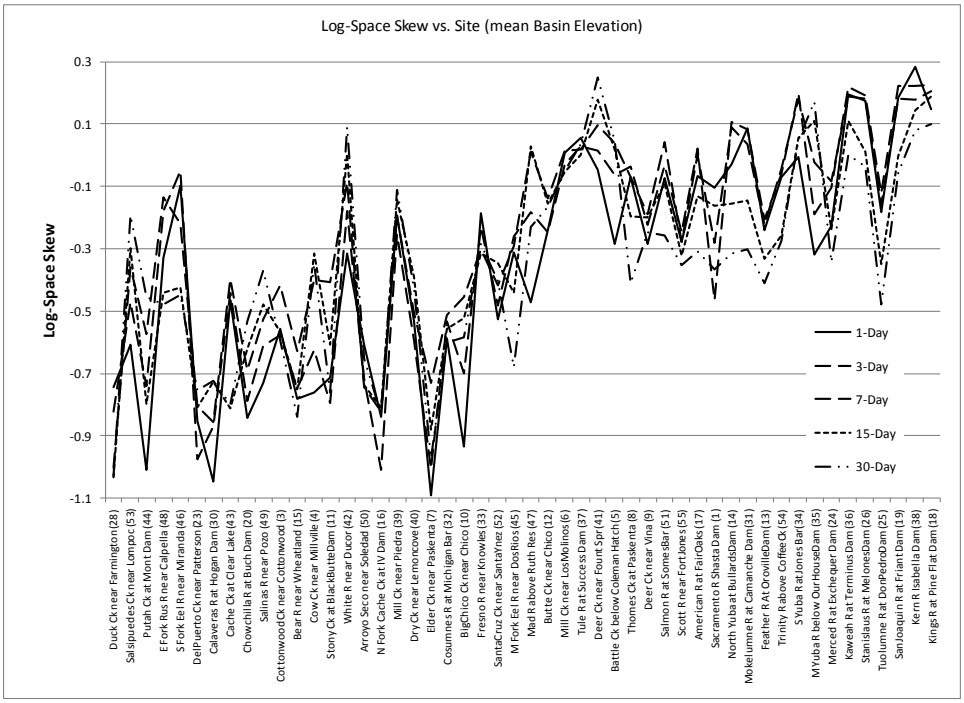


**Figure 4.5:** Observed at-site sample skew coefficients versus site, ascending mean basin elevation.

To address the apparent non-linear trend in elevation, a variety of non-linear functions of elevation were tested. Ultimately, the function given by Equation (4.3) provided the best regression results across all durations. Section 4.6 Development procedure for non-linear models and sensitivity analysis describes the development of the non-linear elevation function as well as an assessment of the possible error incurred by ignoring estimation error in its coefficients.

$$NL = 1 - \exp\left[-\left(\frac{Elev}{3600}\right)^{12}\right] \tag{4.3}$$

where $Elev$ is the mean basin elevation in feet.

This function is essentially zero for low elevation basins ($Elev$ less than 2,500 ft) and one for high elevation basins ($Elev$ greater than 4,500 ft), with a rapid transition period between 3,000 and 4,000 ft. This function is similar to that used by Parrett et al. [2011] for California annual maximums.

Many regional skew model forms and explanatory variables were tested. For brevity, only four model forms are presented and discussed in this thesis: the constant model, the linear elevation model, a discontinuous EL6000 model, and the non-linear elevation model.

The constant model has the form:

$$\gamma = \beta_0 \tag{4.4}$$

where

$\beta_0$ is a regression constant

The linear elevation model has the form:

$$\gamma_i = \beta_0 + \beta_1 Elev_i \tag{4.5}$$

where

$\beta_0$ and $\beta_1$ are regression constants
$Elev_i$ is the mean basin elevation for site $i$.

The discontinuous $EL6000$ model has the form:

$$\gamma_i = \beta_0 \quad \text{for } EL6000_i \leq 4$$

and (4.6)

$$\gamma_i = \beta_0 + \beta_2 \quad \text{for } EL6000_i > 4$$

where

$\beta_0$ and $\beta_2$ are regression constants
$EL6000_i$ is the $EL6000$ (percent of basin above 6,000 ft) for site $i$.

The non-linear (NL) elevation model has the form:

$$\gamma_i = \beta_0 + \beta_3 NL_i \tag{4.7}$$

where

$\beta_0$ and $\beta_3$ are regression constants
$NL_i$ is the non-linear ($NL$) function defined in Equation 4.2 for site $i$.

The discontinuous $EL6000$ model was developed because a linear $EL6000$ model failed to adequately describe basins with low $EL6000$. As an alternative, the three parameter model described by Equation (4.7) was developed:

$$\gamma_i = \beta_0 \qquad \text{for } EL6000_i \leq 4$$

and (4.8)

$$\gamma_i = \beta_0 + \beta_2 + \beta_4 EL6000_i \qquad \text{for } EL6000_i > 4$$

where

$\beta_0$, $\beta_2$, and $\beta_4$ are regression constants
$EL6000_i$ is the $EL6000$ for site $i$

The slope coefficient $\beta_4$ failed to be statistically significant for any duration, so the model described by Equation 4.6 was adopted instead.

**Table 4.3:** Summary of statistical results for various models considered. Terms in parenthesis are standard error of computed term above.

| Duration | Type | B0 | B1 | B2 | B3 | MEV | ASVE | $AVP_{new}$ | $R^2$ | Nominal ERL |
|---|---|---|---|---|---|---|---|---|---|---|
| | Constant* | **-0.32** | - | - | - | 0.078 | 0.035 | 0.113 | 0 | 66 |
| **1-Day** | Linear Elevation | -1.02 | 1.79E-04 | - | - | 0.026 | 0.040 | 0.066 | 0.66 | 110 |
| | Discont. EL6000 | -0.69 | - | 0.62 | - | 0.017 | 0.038 | 0.055 | 0.78 | 131 |
| | NL Elevation | -0.73 | - | - | 0.69 | 0.012 | 0.038 | 0.049 | 0.85 | 146 |
| | NL Elev Final** | -0.73 (0.22) | - | - | 0.69 (0.12) | 0.011 (0.009) | 0.037 | 0.048 | 0.86 | 150 |
| | Constant | **-0.27** | - | - | - | 0.080 | 0.039 | 0.118 | 0 | 62 |
| **3-Day** | Linear Elevation | -0.97 | 1.78E-04 | - | - | 0.025 | 0.043 | 0.068 | 0.69 | 104 |
| | Discont. EL6000 | -0.64 | - | 0.63 | - | 0.016 | 0.041 | 0.057 | 0.80 | 122 |
| | NL Elevation | -0.68 | - | - | 0.71 | 0.008 | 0.040 | 0.049 | 0.90 | 143 |
| | NL Elev Final | -0.69 (0.22) | - | - | 0.68 (0.11) | 0.009 (0.008) | 0.040 | 0.049 | 0.89 | 143 |
| | Constant | **-0.22** | - | - | - | 0.053 | 0.040 | 0.093 | 0 | 76 |
| **7-Day** | Linear Elevation | -0.83 | 1.53E-04 | - | - | 0.014 | 0.045 | 0.059 | 0.74 | 117 |
| | Discont. EL6000 | -0.54 | - | 0.53 | - | 0.013 | 0.043 | 0.056 | 0.75 | 121 |
| | NL Elevation | -0.58 | - | - | 0.61 | 0.007 | 0.042 | 0.049 | 0.87 | 138 |
| | NL Elev Final** | -0.59 (0.23) | - | - | 0.59 (0.11) | 0.007 (0.006) | 0.042 | 0.049 | 0.87 | 140 |
| | Constant | **-0.30** | - | - | - | 0.034 | 0.043 | 0.076 | 0 | 95 |
| **15-Day** | Linear Elevation | -0.88 | 1.44E-04 | - | - | 0.010 | 0.048 | 0.058 | 0.71 | 124 |
| | Discont. EL6000 | -0.60 | - | 0.49 | - | 0.008 | 0.046 | 0.055 | 0.75 | 130 |
| | NL Elevation | -0.65 | - | - | 0.57 | 0.006 | 0.046 | 0.052 | 0.84 | 138 |
| | NL Elev Final** | -0.65 (0.24) | - | - | 0.55 (0.11) | 0.005 (0.005) | 0.046 | 0.051 | 0.85 | 141 |
| | Constant | **-0.36** | - | - | - | 0.033 | 0.044 | 0.076 | 0 | 98 |
| **30-Day** | Linear Elevation | -0.84 | 1.21E-04 | - | - | 0.017 | 0.049 | 0.066 | 0.48 | 113 |
| | Discont. EL6000 | -0.60 | - | 0.40 | - | 0.012 | 0.047 | 0.059 | 0.63 | 125 |
| | NL Elevation | -0.64 | - | - | 0.47 | 0.011 | 0.047 | 0.058 | 0.67 | 128 |
| | NL Elev Final** | -0.63 (0.24) | - | - | 0.44 (0.11) | 0.010 (0.008) | 0.046 | 0.056 | 0.69 | 133 |

* Non-significant terms in bold

** Analysis included all 50 sites, whereas all others included 47 sites

Table 4.3 provides a summary of the statistical analysis for the best performing models for all durations. The diagnostic and goodness of fit statistics are defined and described in Section 3.2. Nominal effective record length (nominal ERL) is a variation of the previously defined ERL. Since the variance of prediction and the regional skew coefficient varies across the range of the explanatory variables, it is impossible to give a single ERL for a skew model unless it is constant. Nominal ERL instead uses the average variance of prediction for a new site ($AVP_{new}$) and the constant model regional skew coefficient to calculate the ERL. The actual ERL for a new or old basin depends on the mean elevation of that basin.

The constant model coefficient fails to be statistically different than zero at the 5% level for any duration considered in this study, indicating that there is not strong evidence that the coefficient should not be zero. This is surprising since the majority of the sample skew coefficients are negative, and the model error variance (MEV) is relatively small. This is explained by the high cross-correlation between sample skew coefficients, which greatly affects the precision of the estimated parameters [Stedinger and Tasker, 1985]. The misrepresentation of beta variance (MBV) ranged between 13.4 and 18.4 across study durations for the final models, indicating that ignoring the cross-correlation would cause the analysis to overestimate the precision of the constant term in Equation (4.7) by a factor of 13.4 to 18.4. While it is not a perfect analog for the constant model, it gives an idea of the effect of cross-correlation on model precision.

Cross-correlation reduces the precision of the regression parameters by reducing the effective number of independent sites. Stedinger [1983, eqn. 41] gives an

approximation for the variance of a regional average skew given $K$ sites, each with $n$ normally distributed observations, and average cross-correlation of $\bar{\rho}$. With this approximation, it is possible to compute the effective number of independent sites for any combination of $n$, $K$, and $\bar{\rho}$. Figure 4.6 plots the effective number of independent sites versus average cross-correlation for the case that $K$ is 50, and $n$ is 57, which is the average years of concurrent record in this study.

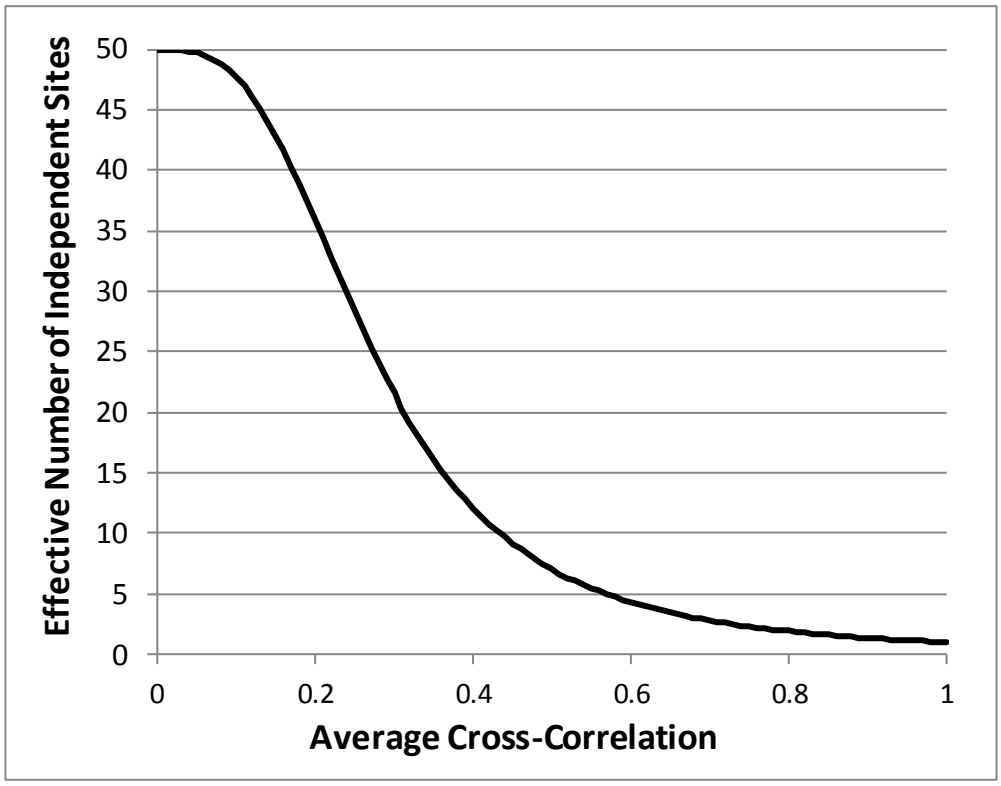


**Figure 4.6:** Effective Number of Independent Sites vs. Average Cross-Correlation, using Stedinger [1983] approximation

The average cross-correlation in this study is 0.283, though this varies across sample sites. The effective number of independent sites for the case that $\bar{\rho}$ is 0.283 is about 23: less than half of the actual number of sites included. Of course this is an approximation, some of whose assumptions are violated, but it gives a sense for the

affect of cross-correlation on model precision. The impact of this effect on the Pseudo ANOVA table (Table 4.4) is explored more in Section 5.2.

The $R_\delta^2$ describes the amount of variation in the at-site sample skew coefficients the model is describing, so the $R_\delta^2$ for the constant model is zero for all durations. The linear elevation model performs well, with $R_\delta^2$ between 0.48 and 0.74, and nominal ERL lengths between 104 and 124 years depending on duration. However, the linear model tends to over-estimate skew coefficients for low elevation basins and under-estimate skew coefficients for high elevation basins. The discontinuous $EL6000$ model performs better than the linear elevation model for all durations, with $R_\delta^2$ between 0.63 and 0.80 and nominal ERL lengths between 104 and 131 years depending on duration. The non-linear elevation model performs best of all the models tested, with $R_\delta^2$ between 0.67 and 0.90 and nominal ERL lengths between 128 and 146 years depending on duration.

The discontinuity of the $EL6000$ model presents some difficulty. As formulated in Equation (4.6), the model assigns one regional skew coefficient to low $EL6000$ basins and another to high $EL6000$ basins. The discontinuity is not a concern for most of the range of $EL6000$, but might present a problem for basins with $EL6000$ near 4. In this case, a river basin evaluated at one point along the channel might be given a regional skew coefficient value vastly different than at a second nearby upstream point, despite the basins being almost completely redundant. In this case, it seems unlikely that flood characteristics at one site are very different than those at another site. The non-linear elevation model addresses this by essentially

providing one regional skew value for low elevation basins and another for high elevation basins, with a rapid, smooth, and continuous transition in between.

Figures Figure 4.7 through Figure 4.11 plot the observed at-site sample skew coefficients and the fitted non-linear elevation regional skew model versus elevation for each duration considered in this study. Figure 4.12 compares the five fitted models. The wide scatter exhibited in the sample skew coefficients displayed in Figures Figure 4.7-Figure 4.11 is mostly due to the sampling error in the skew coefficient estimators. Moreover, because of the high correlations among the annual peaks of a given duration, the residual errors are correlated. For example, the cluster of three high-elevation sites with very positive skew observed in Figure 4.11 correspond to the San Joaquin River (study basin 19), the Kings River (study basin 18), and the Kern River (study basin 38), which are adjacent to each other in the highest region of the Sierra Nevada Mountain Range.

Overall the model errors are remarkably small, and the fitted functions are reasonable (see Table 4.4). For the rainfall flood series considered here, there is a clear distinction between basins with low mean elevation and basins with high mean elevation. Lower mean elevation basins are likely completely dominated by rainfall only floods, while higher mean elevation basins experience rain-on-snow events which make very small annual maximums much less likely, and change the shape of the flood distribution.

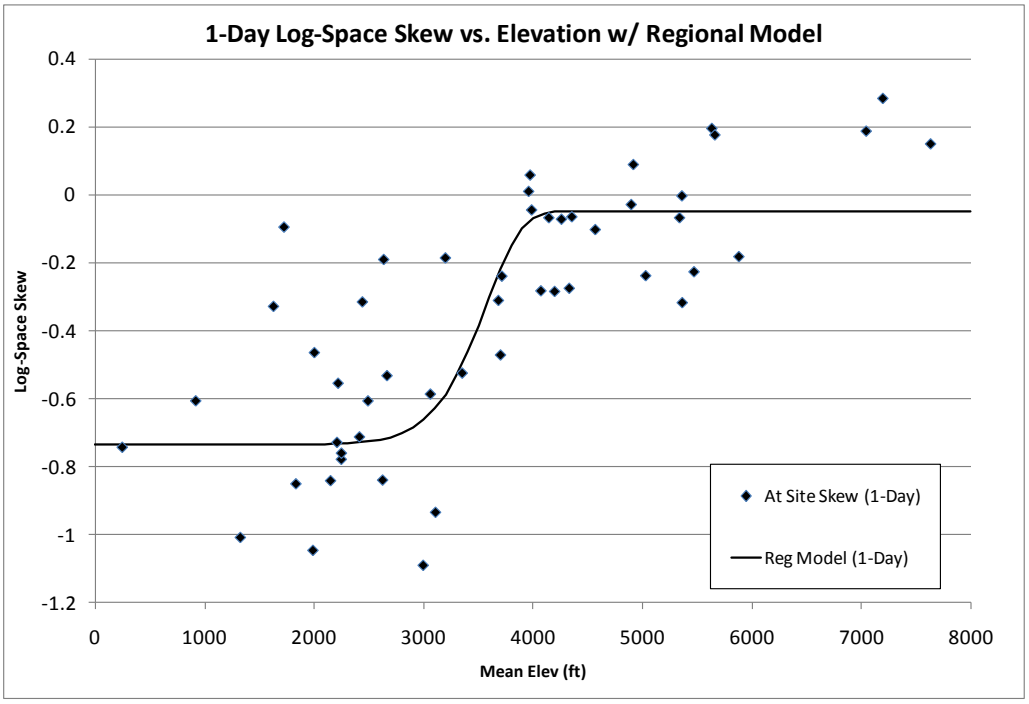


**Figure 4.7:** Observed at-site sample skew coefficients versus mean basin elevation (1-day).

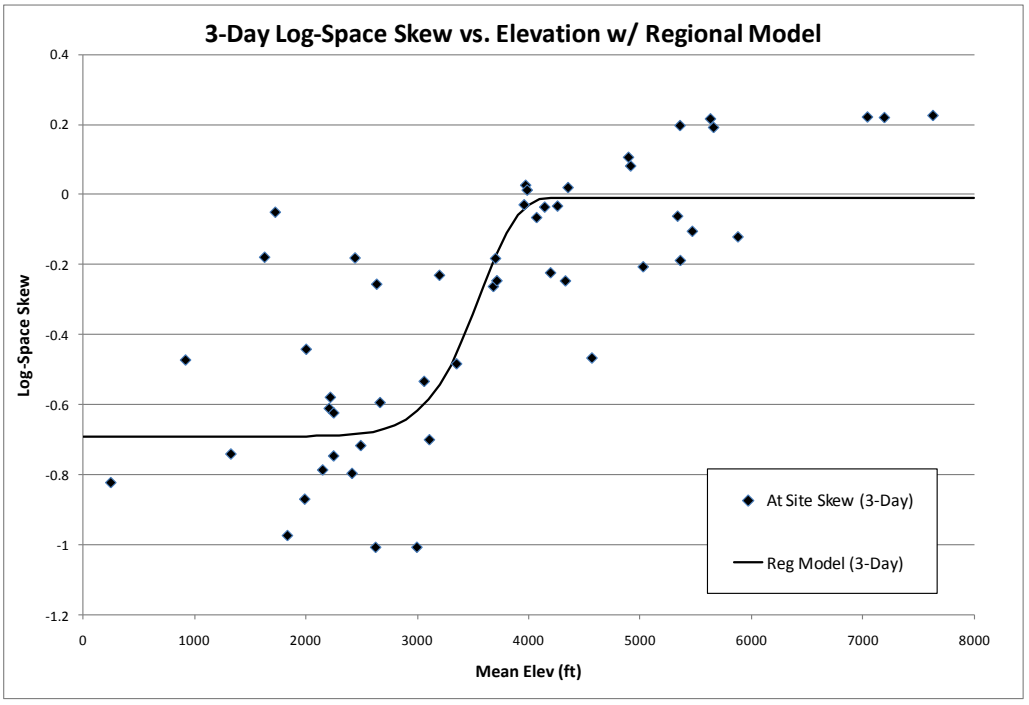


**Figure 4.8:** Observed at-site sample skew coefficients versus mean basin elevation (3-day).

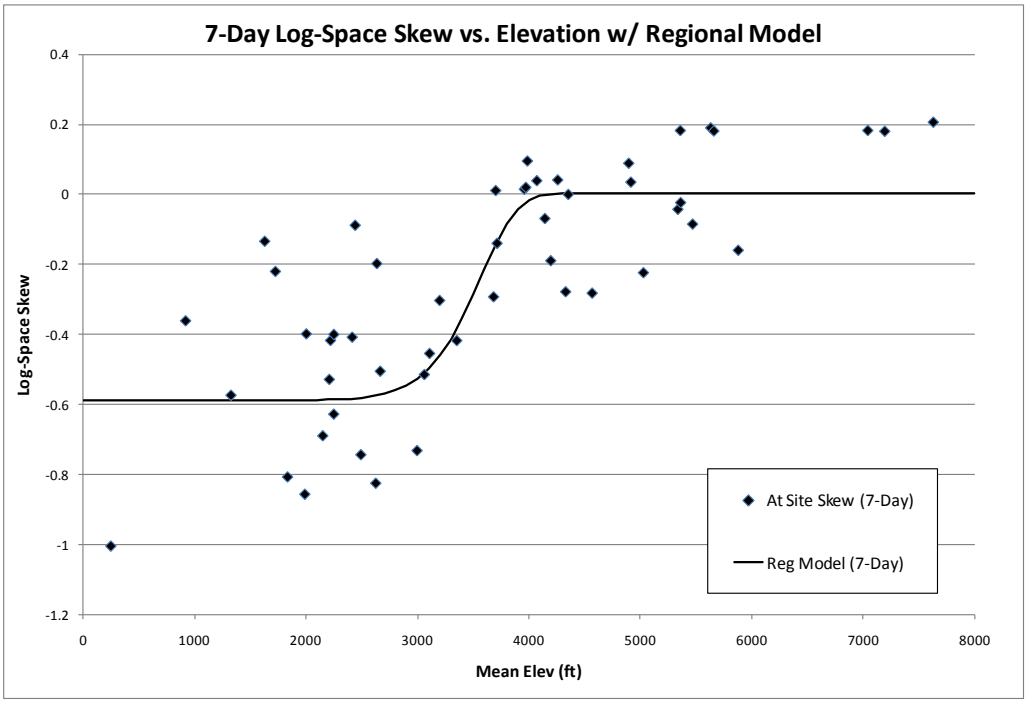


**Figure 4.9:** Observed at-site sample skew coefficients versus mean basin elevation (7-day).

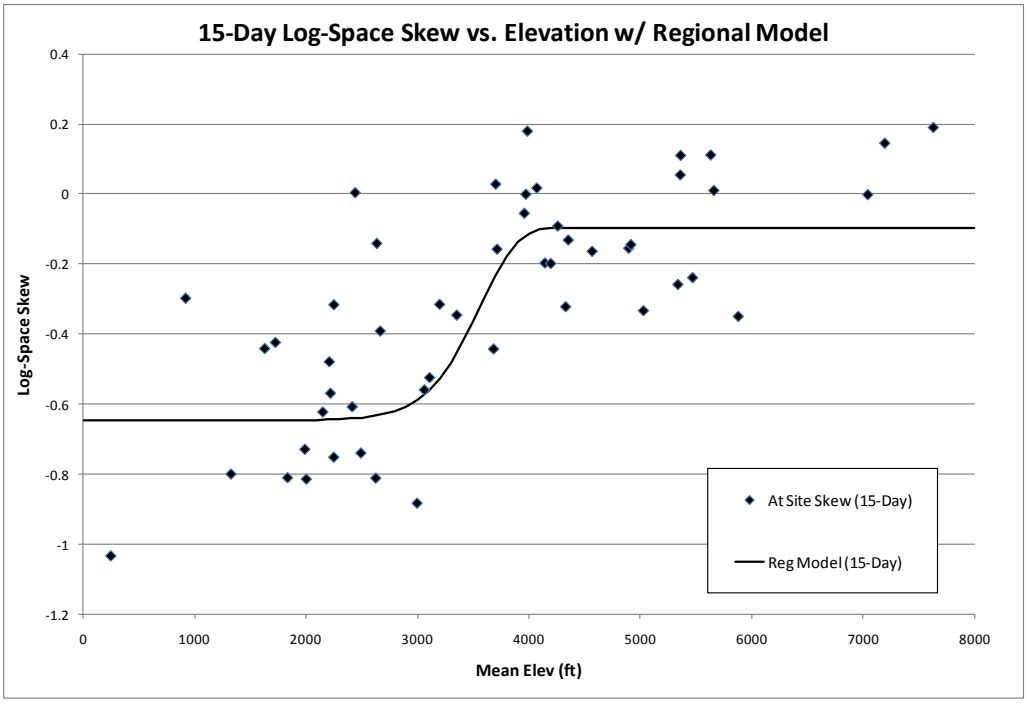


**Figure 4.10:** Observed at-site sample skew coefficients versus mean basin elevation (15-day).

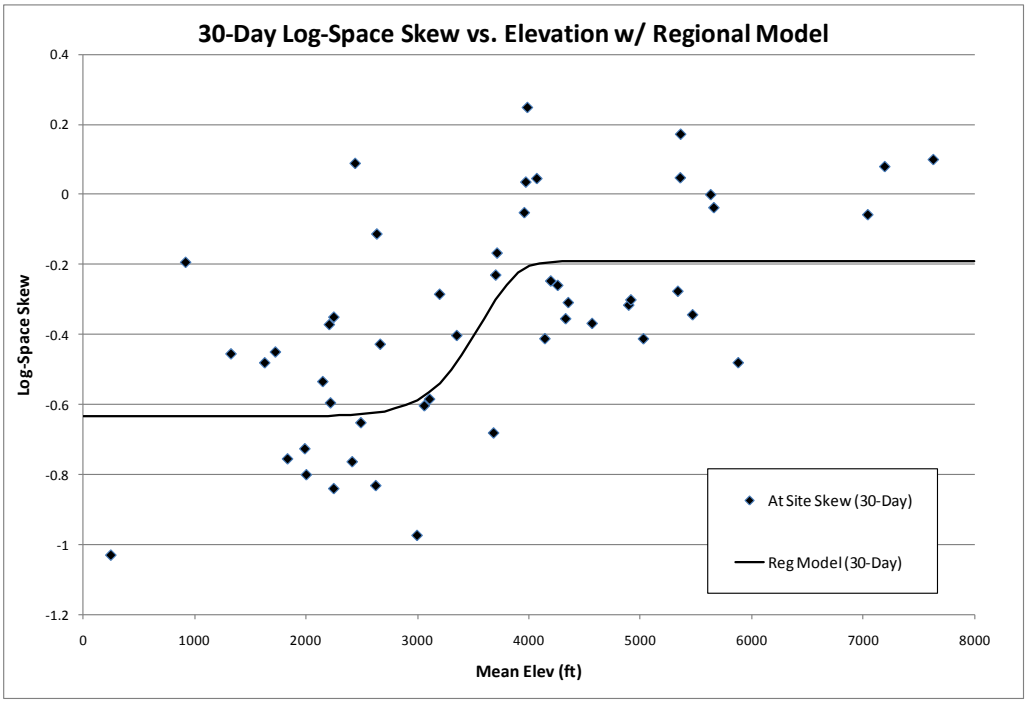


**Figure 4.11:** Observed at-site sample skew coefficients versus mean basin elevation (30-day).

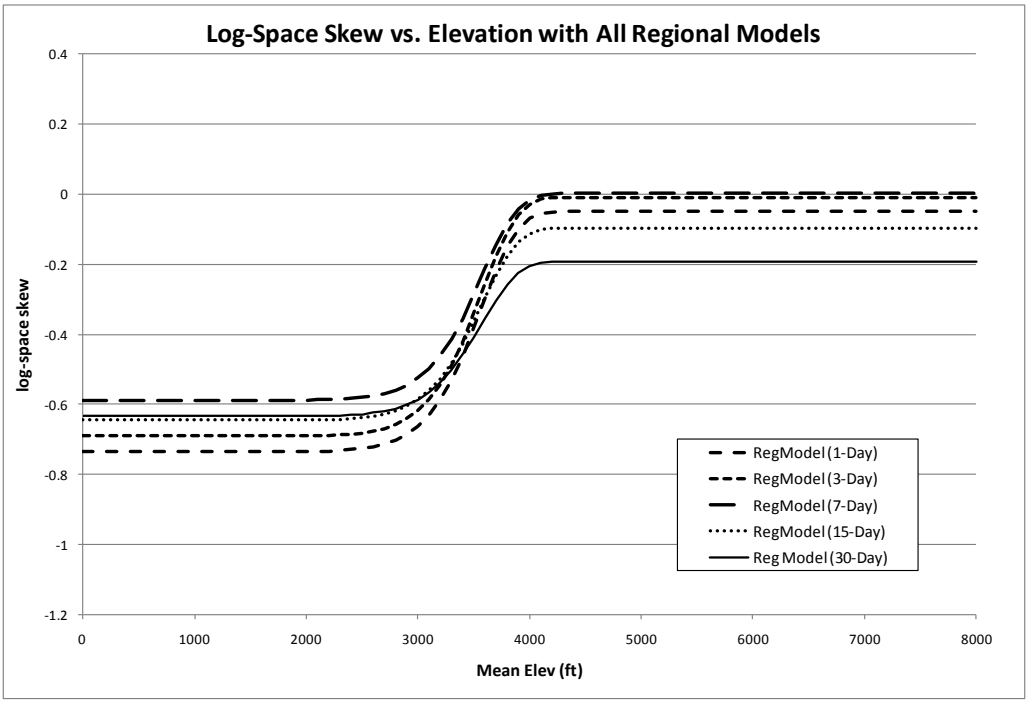


**Figure 4.12:** Regional Skew Models for all durations considered.

Figure 4.12 plots the fitted regional skew coefficient model for each of the durations considered in this study. It should be noted that the 30-day skew model diverges somewhat from the other models and that the models are not ordered as one might expect. It is helpful to consider that these are skew coefficient models, and not flow quantile models, so they should not necessarily be sorted by duration. In fact, if one considers the standard errors of the model parameters (roughly 0.2 for the constant and 0.1 for slope), it is not even clear that the models are significantly different from each other. Thus, hydrologic insight based on differences between models must be stated cautiously and discussed with some skepticism.

Some of the trends exhibited do seem to reinforce our hydrologic understanding of rainfall floods in California. The hydrologic driver of the 30-day duration is clearly a different process than the hydrologic drivers of the shorter durations. The 1, 3, and 7-day duration events are almost always caused by a single passing storm system, whereas the 15-day and 30-day duration events are almost certainly caused by several weeks of storm systems. Thus it is not surprising that their flood distribution characteristics are different. It can be observed that the skew models for shorter duration events (1, 3, and 7-day) are in fact sorted by duration and do not cross; the longer the flood duration, the less negative the skew coefficient at all elevations.

The 15-day and 30-day duration skew coefficient models vary much less between high and low elevation basins, indicating that at longer durations, flood distribution characteristics are more uniform across the study region than at shorter durations. The averaging of flood volumes over longer durations is at least in part

responsible for this increased uniformity with increased duration. Another possible explanation is that basin response to prolonged rain events (two to four weeks) is more uniform because the rain volume exceeds the natural basin storage quickly relative to the flood duration. This minimizes the impact of basin heterogeneity on flood characteristics because a greater percent of the total rainfall is going directly to runoff. Given this, it is not surprising that much more scatter was observed in the sample skew coefficients for 30-day duration than shorter durations (see Figures Figure 4.7-Figure 4.11), and also that less of an elevation signal was observed (see "Model" term Table 4.4 and Pseudo $R_{\delta}^{2}$ in Table 4.3).

Table 4.4 reports a pseudo ANalysis Of VAriance (ANOVA) for the recommended non-linear model for each duration. The table divides the variability observed in the skew coefficient estimators into three categories: the variability explained by the model ("Model"), the variability in the true skews that the model cannot explain ("Model Error"), and the variability due to the sampling error in the individual skew coefficient estimators ("Sampling Error"). The table also contains the total variability, which is the sum of the three. The major source of variability for all durations is the sampling error. Recall from Section 3.2, the Error Variance Ratio (EVR) is the average sampling variance divided by the variance of the model error. For this study, EVR values range from 12.4 to 26.3 across durations. Thus, the sampling error in the skew coefficient estimators is overall much larger than the model error. Clearly an Ordinary Least Squares analysis, which ignores the limited precision of the skew coefficient estimators, is likely to fail to correctly represent the information in the data set and ultimately the true variance of prediction.

**Table 4.4:** Pseudo ANOVA for fitted model for each duration considered

| | 1-Day | 3-Day | 7-Day | 15-Day | 30-Day |
|---|---|---|---|---|---|
| **Model** | 3.384 | 3.660 | 2.347 | 1.359 | 1.070 |
| **Model Error** | 0.533 | 0.448 | 0.336 | 0.244 | 0.485 |
| **Sampling Error** | 6.602 | 6.439 | 6.234 | 6.399 | 6.348 |
| **Total** | 10.519 | 10.548 | 8.916 | 8.002 | 7.902 |
| | | | | | |
| **EVR** | 12.4 | 14.4 | 18.6 | 26.3 | 13.1 |
| **MBV** | 13.4 | 15.2 | 17.1 | 18.4 | 18.0 |
| **Pseudo $R_\delta^2$** | 0.86 | 0.89 | 0.87 | 0.85 | 0.69 |

Table 4.4 also reports the Misrepresentation of Beta Variance (MBV) statistic, which is the ratio of the sampling variance that the GLS analysis ascribes to the constant in the model and the variance that a WLS analysis (which neglects cross-correlations) would ascribe to the constant (see Section 3.2) [Parrett et al., 2011]. MBV values range from 13.4 to 18.4 across the durations. These large values of MBV indicate that error analysis produced by WLS would drastically overestimate the precision of the constant term. Thus, a GLS analysis is needed to correctly evaluate the precision with which the constant term can be resolved. This is particularly important in these analyses because the contribution of parameter uncertainty to the average variance of prediction is at least twice as large as the model error variance for each study duration. This is a result of the very large correlation among the records which limits the amount of information the regional data set provides.

The total variability decreases with increasing duration. This means that as duration increases there is less elevation signal in the observed skew coefficients and that skew coefficient values become more uniform. This is confirmed by observing Figures 4.6A-4.6E: note that as duration increases, so too does the scatter in skew

values. Not surprisingly, the amount of variability the model is able to describe also decreases with increasing duration, as there is simply less variability to explain. Model error remains relatively constant across durations, as does sampling error.

Because the models include an explanatory variable that depends on elevation, the actual variance of prediction for a site depends on its average basin elevation. Table 4.5 gives the variance of prediction for a new site (a site not included in this study) as a function of its mean basin elevation between 0 and 10,000 feet. For basins below 2,500 feet, there is no change in the variance of prediction with elevation; similarly, all basins above 4,500 feet have the same variance of prediction. There is little variation in the effective record lengths (ERL) with changes in elevation despite an appreciable variation in the variance of prediction. The change in the sampling variance of the skew coefficient estimators due to the change in estimated skew coefficient with elevation approximately balanced the differences in the prediction variance for lower and higher elevations.

**Table 4.5:** Variance of Prediction ($VP_{new}$) and Effective Record Length (ERL) for all durations as a function of elevation.

| Elevation (ft) | 1-Day | | 3-Day | | 7-Day | | 15-Day | | 30-Day | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $VP_{new}$ | ERL | $VP_{new}$ | ERL | $VP_{new}$ | ERL | $VP_{new}$ | ERL | $VP_{new}$ | ERL |
| < 2500 | 0.058 | 186 | 0.059 | 172 | 0.058 | 156 | 0.062 | 157 | 0.066 | 145 |
| 3000 | 0.055 | 182 | 0.056 | 168 | 0.055 | 155 | 0.059 | 156 | 0.063 | 144 |
| 3200 | 0.052 | 177 | 0.053 | 164 | 0.053 | 153 | 0.055 | 155 | 0.060 | 144 |
| 3400 | 0.047 | 170 | 0.049 | 159 | 0.049 | 151 | 0.051 | 154 | 0.056 | 143 |
| 3600 | 0.043 | 164 | 0.044 | 155 | 0.045 | 151 | 0.046 | 154 | 0.052 | 142 |
| 3800 | 0.040 | 162 | 0.042 | 155 | 0.042 | 153 | 0.042 | 156 | 0.049 | 141 |
| 4000 | 0.039 | 162 | 0.041 | 157 | 0.041 | 156 | 0.041 | 157 | 0.048 | 141 |
| >4500 | 0.039 | 162 | 0.040 | 157 | 0.041 | 156 | 0.041 | 157 | 0.048 | 140 |

***Section 4.5 Leverage and Influence measures***

Leverage can be thought of as the potential impact of an observation on the fitted model due to its position in the variable space. Influence is a measure of the actual effect of an observation on the final fitted regression model coefficients. Leverage and influence terms are formally defined in Section 3.2, and Section 3.2.2 derives the new leverage and influence values used in this study.

If $\hat{\boldsymbol{\beta}}$ has dimension *k* and *n* is the sample size (number of basins in the study), the mean of the leverage values is *k/n* and values greater than *2k/n* are generally considered large. Influence values greater than 4/*n* are typically considered large [Tasker and Stedinger, 1989]. Using these guidelines, leverage values greater than 0.12 and influence values greater than 0.08 were considered large in this study.

Figures Figure 4.13-Figure 4.17 display influence and leverage statistics for each basin for the 1-day, 3-day, 7-day, 15-day, and 30-day durations sorted by influence, respectively. Leverage values did not change radically from one duration to another because the matrix of basin characteristics and the sample sizes were the same for all durations. The small changes in leverage values are because the at-site skew coefficients for a basin were different for different durations as were the model error variances. Because influence values depend on the residuals computed from the individual skew estimators for each duration, there was significant variation in the influence of some basins from one duration to another.

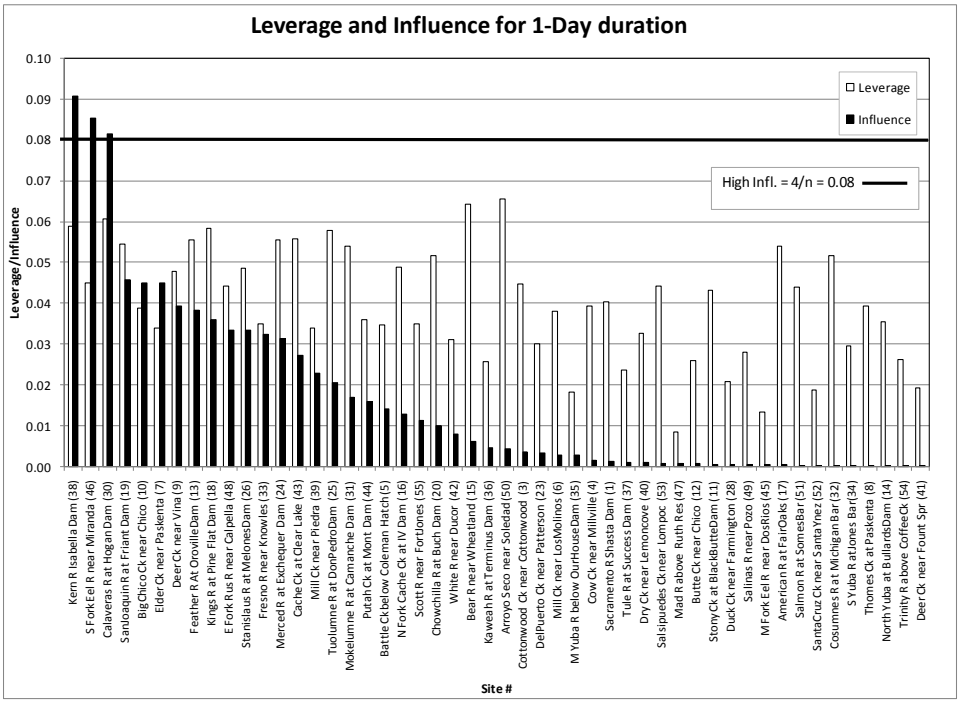


**Figure 4.13:** Leverage and Influence values (1-day), sorted by influence.

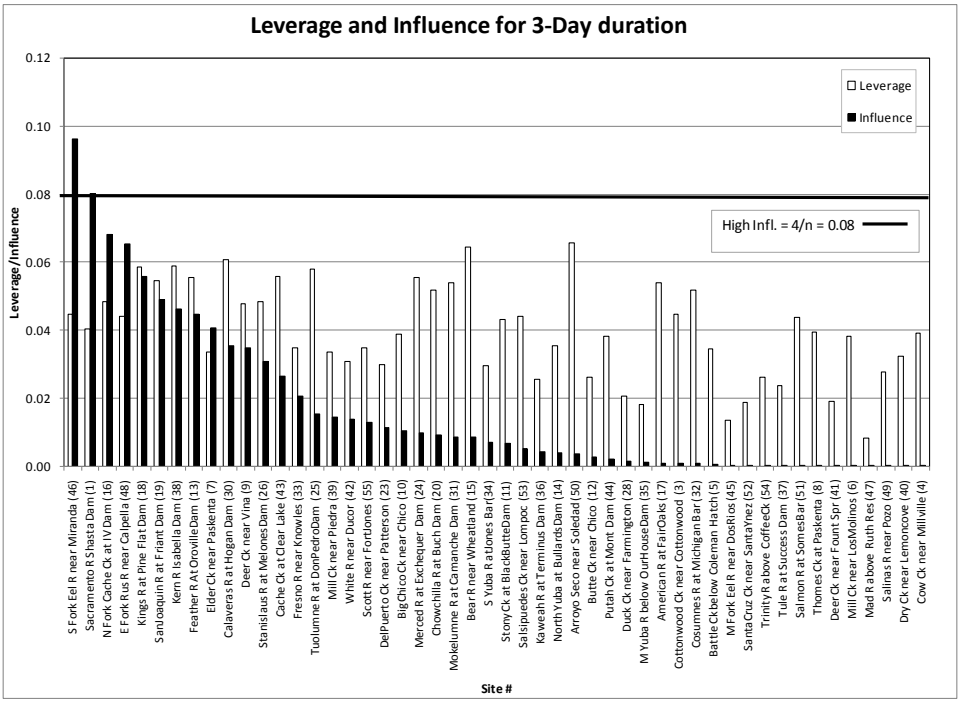


**Figure 4.14:** Leverage and Influence values (3-day), sorted by influence.

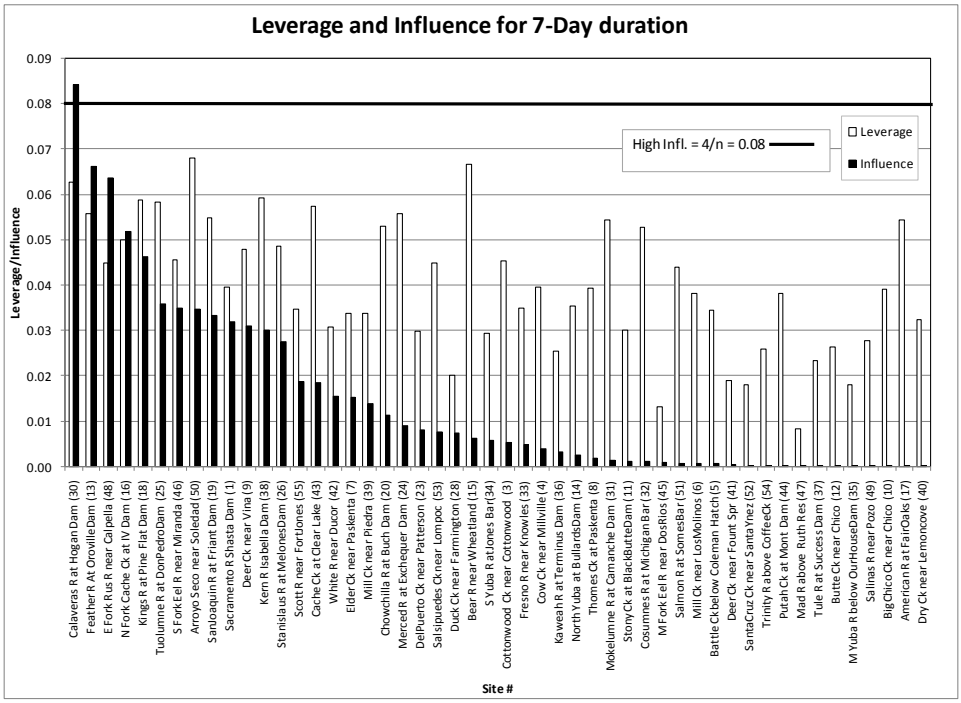


**Figure 4.15:** Leverage and Influence values (7-day), sorted by influence.

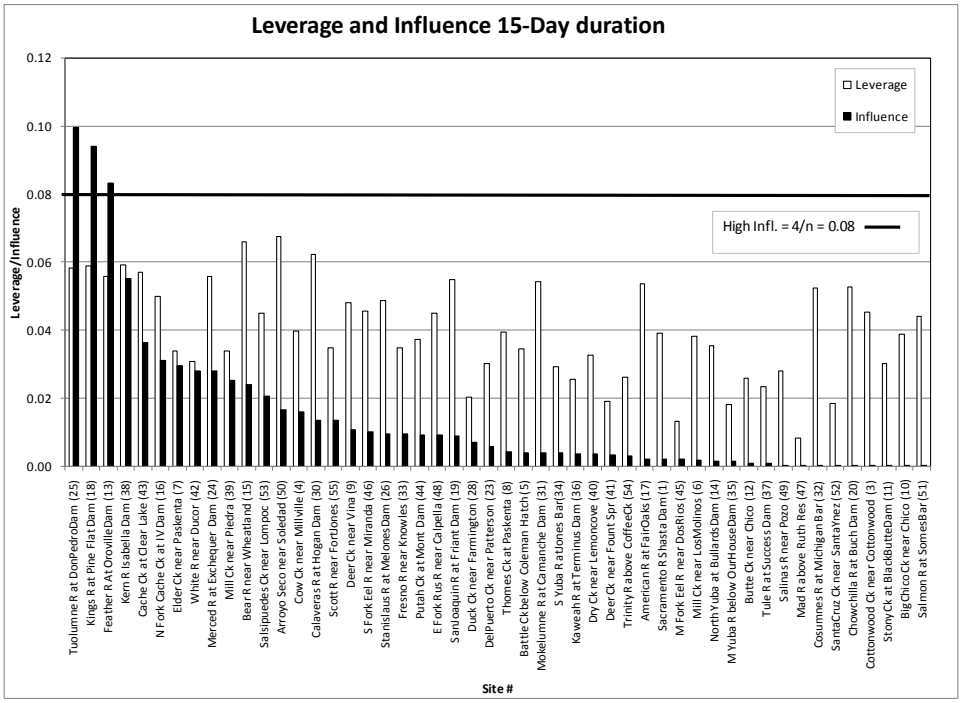


**Figure 4.16:** Leverage and Influence values (15-day), sorted by influence.

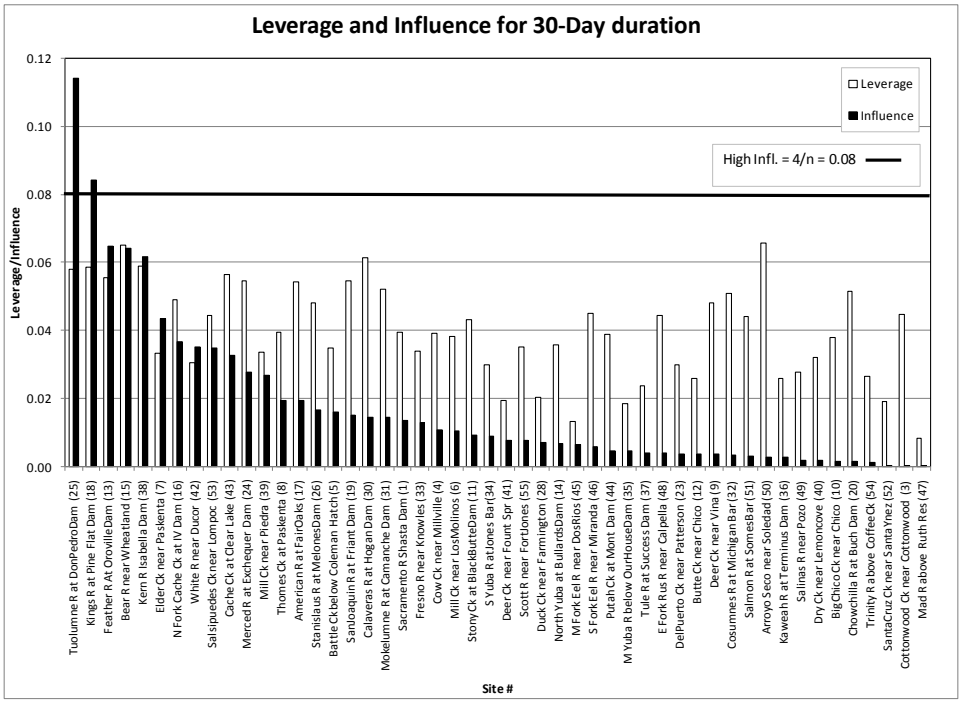


**Figure 4.17:** Leverage and Influence values (30-day), sorted by influence.

The South Fork Eel River near Miranda (study basin 46) has high influence at shorter durations. This is because the South Fork Eel River had a very high residual for the 1-day and 3-day durations, with a relatively small sampling error variance (based on 68 years of record). The influence, particularly for 1-day, was only marginally high and therefore not important to the study.

The Tuolumne River at New Don Pedro Dam (study basin 25) and the Kings River at Pine Flat Dam (study basin 18) have high influence at longer durations. Each of these basins has a long record length (112 and 113 years, respectively), resulting in larger weights and thus relatively high leverage. Since they have large residuals at the longer durations and long record lengths, it was not surprising their influence values are so high.

Upon first inspection of the leverage and influence values for this study, the Sacramento River at Shasta Dam (study basin 1) appeared to have very high influence for the 1-day duration, but only moderate to low influence for all other durations. This warranted a closer examination of the 1-day data for this basin, which revealed an error in the censoring level: a clear low outlier had not been censored, resulting in an uncharacteristic highly negative skew. After this additional value was censored and the regression was re-run, the regional skew model fit was improved, and the Sacramento River basin's influence for the 1-day duration became quite reasonable. Generally, it is poor practice to change the number of observations censored explicitly to achieve a desired result, but in this case, the diagnostic statistics alerted the researcher to an error in a previous analysis, which was corrected and resulted in an improved model.

Overall, this is an example wherein large leverage values were not expected. The value of the nonlinear function of elevation ranged from zero for basins below 3,000 feet to 1 for basins above 4,200 feet. Thus, it was impossible for any basin to have an extreme value. Sampling error associated with each skew coefficient also contributed to the leverage. Longer-record sites did not have record lengths much longer than 100 years, and many sites had record lengths about that long. Thus, no sites were unusual. Examining the leverage and influence statistics indicated there were no obvious problems in the development of the flood data, the basin characteristics file, the at-site skew estimators, or in the statistical analyses.

***Section 4.6 Development procedure for non-linear models and sensitivity analysis***

As exhibited in Figure 4.12, the observed at-site skew coefficients in this study followed an apparent non-linear trend in mean basin elevation. A similar phenomenon was observed by the California instantaneous annual maximum skew study. The effect was attributed to a transition from rain-only floods in low elevation basins to rain-on-snow floods in high elevation basins [Parrett et al., 2011, pp. 15].

Parrett et al. [2011] experimented with a variety of nonlinear functions of elevation. The selected function NL, as described by Equation (4.9) varies between zero at low elevations to one at high elevations.

$$NL = 1 - \exp\left(-\left(\frac{Elev}{a}\right)^{b}\right) \tag{4.9}$$

Here $a$ is a scale parameter that determines the location of the transition between high and low basins, and $b$ is a slope parameter which determines how rapidly NL increases from 0 and 1 near $a$.

Parrett et al. [2011] found that an NL term with $a = 6500$ ft and $b = 2$ described a trend from low to high elevation basin skew coefficients well when scaled and added to a constant term. This study adopted the same non-linear form, but found that the scale and slope factors used by the previous study were not the best for rainfall floods. In particular, it appeared that the step between high and low elevation basin skews occurred at a lower elevation than previously observed. Also, the zone between low and high elevation basin skews was much shorter and the rise more dramatic than in the previous study of annual instantaneous maximum floods. Regional skew models utilizing several NL terms with various combinations of $a$ and $b$ such that

$a \leq 6500$ and $b \geq 2$ were fit for each study duration. The NL term with a location parameter *a* of 3,600 ft and a slope parameter *b* of 12 provided a good fit to the observed skew coefficients across all durations.

Figure 4.18 plots the NL term used in this study (a = 3600, b = 12) and the NL term used by Parrett et al. [2011] (a = 6500, b = 2) versus mean basin elevation. The NL(a = 6500, b = 2) term rises much more gradually than NL(a = 3600, b = 12), appearing almost linear over the range of mean basin elevation shown in Figure 4.18. In contrast, the rise of NL(a = 3600, b = 12) is so rapid that NL becomes either 0 or 1 for most study basins.

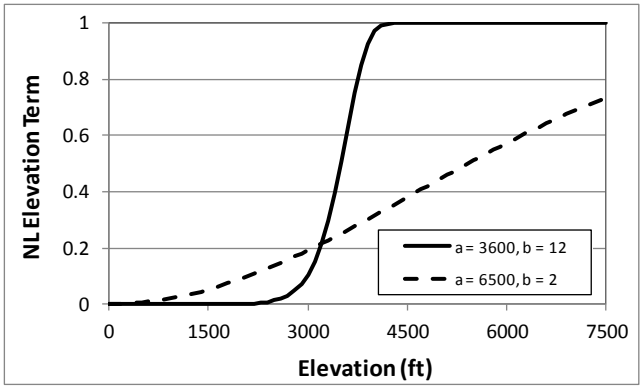


**Figure 4.18:** Non-linear Elevation Term (NL) versus mean basin elevation for this study (a = 3600, b = 12) and the California instantaneous maximum study (a = 6500, b = 2).

A summary of statistical results for various parameterizations of NL are reported in Table 4.6. Other parameterizations were also tested, but the models in Table 4.6 resulted in the best fits among those tested.

**Table 4.6:** Summary of Bayesian WLS/GLS statistical results for various non-linear models considered in the California Rainfall flood Skew Study.

| Duration | Type | B0 | B1 | MEV | ASVE | $AVP_{new}$ | $R^2$ | Nominal ERL |
|---|---|---|---|---|---|---|---|---|
| **1-Day** | a = 4000 b = 4 | -0.80 | 0.86 | 0.015 | 0.038 | 0.053 | 0.81 | 136 |
| | a = 4000 b = 8 | -0.71 | 0.74 | 0.013 | 0.038 | 0.051 | 0.83 | 141 |
| | a = 3600 b = 12 | -0.73 | 0.69 | 0.011 | 0.037 | 0.048 | 0.86 | 150 |
| | a = 4500 b = 1 | -1.41 | 1.95 | 0.024 | 0.039 | 0.064 | 0.70 | 114 |
| **3-Day** | a = 4000 b = 4 | -0.75 | 0.85 | 0.016 | 0.041 | 0.057 | 0.81 | 128 |
| | a = 4000 b = 8 | -0.67 | 0.73 | 0.013 | 0.041 | 0.053 | 0.85 | 135 |
| | a = 3600 b = 12 | -0.69 | 0.68 | 0.009 | 0.040 | 0.049 | 0.89 | 146 |
| | a = 4500 b = 1 | -1.34 | 1.91 | 0.029 | 0.043 | 0.072 | 0.65 | 102 |
| **7-Day** | a = 4000 b = 4 | -0.64 | 0.73 | 0.010 | 0.043 | 0.053 | 0.83 | 137 |
| | a = 4000 b = 8 | -0.56 | 0.62 | 0.008 | 0.043 | 0.051 | 0.85 | 142 |
| | a = 3600 b = 12 | -0.59 | 0.59 | 0.007 | 0.042 | 0.049 | 0.87 | 147 |
| | a = 4500 b = 1 | -1.16 | 1.66 | 0.017 | 0.044 | 0.061 | 0.69 | 119 |
| **15-Day** | a = 4000 b = 4 | -0.69 | 0.67 | 0.008 | 0.047 | 0.054 | 0.77 | 133 |
| | a = 4000 b = 8 | -0.61 | 0.57 | 0.007 | 0.046 | 0.053 | 0.80 | 136 |
| | a = 3600 b = 12 | -0.65 | 0.55 | 0.005 | 0.046 | 0.051 | 0.85 | 142 |
| | a = 4500 b = 1 | -1.19 | 1.57 | 0.011 | 0.048 | 0.058 | 0.69 | 125 |
| **30-Day** | a = 4000 b = 4 | -0.67 | 0.54 | 0.015 | 0.047 | 0.062 | 0.56 | 117 |
| | a = 4000 b = 8 | -0.61 | 0.46 | 0.013 | 0.047 | 0.060 | 0.61 | 120 |
| | a = 3600 b = 12 | -0.63 | 0.45 | 0.010 | 0.046 | 0.056 | 0.71 | 128 |
| | a = 4500 b = 1 | -1.07 | 1.26 | 0.019 | 0.048 | 0.068 | 0.43 | 108 |

One concern with the statistical results described in Section 4.4 is that those analyses have assumed the parameters of the NL term in Equation 4.8 were given and not estimated. A particular concern is the use of a universal $a$ parameter for all durations. Consistency between durations was a major concern in the analysis and varying the non-linear term might have led to undesired inconsistencies. On the other hand, the $a$ term might reasonably have varied between 3,000 and 4,000 for each duration. The following analysis considers what would have been the results if a duration specific $a$ had been used rather than a common $a$.

Expanding Equation (4.7), the non-linear elevation model for regional skew at site $i$, $\gamma_i$, can be represented as:

$$\gamma_i = \beta_0 + \beta_3[1 - \exp\left(-\left(\frac{Elev_i}{a}\right)^b\right)] = \beta_0 + \beta_3 - \beta_3 \exp\left(-\left(\frac{Elev_i}{a}\right)^b\right) \tag{4.10}$$

Taking the partial derivative of Equation (4.10) with respect to NL-elevation function parameter $a$, yields:

$$\frac{\partial \gamma_i}{\partial a} = -\beta_3 \exp\left(-\left(\frac{Elev_i}{a}\right)^b\right)\left(\frac{b}{a}\left(\frac{Elev_i}{a}\right)^b\right) \tag{4.11}$$

Equation (4.11) is the change in the regional skew coefficient, $\gamma_i$, for a basin with mean elevation $Elev_i$ given a change in $a$. With this term, a new regional skew model was fit for each duration, having the form:

$$\gamma_i = \beta_0 + \beta_3 NL_i + \beta_4\left(\frac{\partial \gamma}{\partial a} - \overline{\frac{\partial \gamma}{\partial a}}\right) \tag{4.12}$$

where,

$\beta_0$, $\beta_3$ and $\beta_4$ are regression constants
$NL_i$ is the $NL$ function defined in Equation 4.2 for site $i$

$\frac{\partial \gamma_i}{\partial a}$ is given by Equation (4.11)
$\frac{\partial \overline{\gamma}}{\partial a}$ is the sample mean of $\frac{\partial \gamma_i}{\partial a}$ overall study basins

If the parameter $\beta_4$ is not statistically different than zero, it indicates that using a duration specific $a$ results in a model which is not statistically different than the model which assumes a common $a$, i.e. the model in Equation (4.12) is not statistically different than model in Equation (4.10), and use if a universal $a$ is acceptable.

This procedure is similar to the non-linear regression procedure recommended by Draper and Smith [1967, pg. 267]. To apply linear least squares to a non-linear model, they recommend linearization of the non-linear term using a first order Taylor series approximation. Here we have re-added the first order term for the scale parameter to test if a common $a$ for all durations is consistent with the data. Figure 4.19 plots the fitted 1-Day regional skew models described by Equation (4.10) and Equation (4.12).

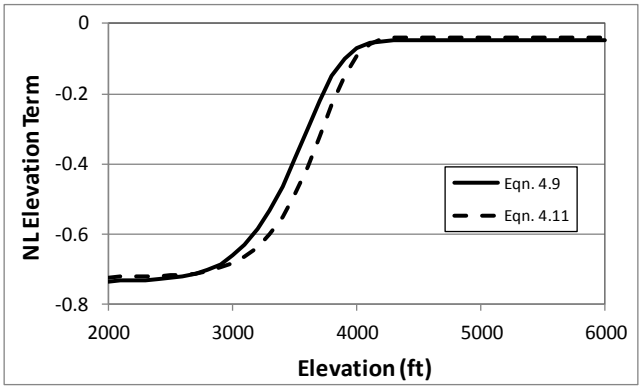


**Figure 4.19:** Fitted 1-day Regional Skew models, using a common non-linear scale parameter, $\boldsymbol{a}$, (Equation (4.10)) and a duration specific $\boldsymbol{a}$ (Equation (4.12**)**).

Table 4.7 contains the fitted regression parameter $\beta_4$ for the model described in Equation (4.12) for each duration, along with its standard error and p-value. Note that $\beta_4$ fails to be statistically significant for any duration. This indicates that neglecting

the first order linearization term is appropriate and that use of a common $a$ parameter across all durations is likely appropriate.

**Table 4.7:** Summary of statistical results for first order linearization term $\boldsymbol{\beta_4}$ of the non-linear elevation model.

| Duration | $\beta_4$ | Std. Error | P-value (Two Sided) |
|---|---|---|---|
| 1 | 141 | 252 | 0.58 |
| 3 | 107 | 250 | 0.67 |
| 7 | 40 | 285 | 0.89 |
| 15 | -94 | 309 | 0.76 |
| 30 | -52 | 390 | 0.89 |

As was resolved earlier, the change in $a$ was not statistically significant. Still each duration specific $a$ would have varied slightly from the universal $a$, and measures of regression precision would change with the introduction of uncertainty in a third parameter. Table 4.8 compares the ERL for the regional skew models described by Equation (4.10) and the expanded model in Equation (4.12) at various elevations. The expanded model serves as a surrogate for estimating a unique $a$ for each duration. Note that ERL varies only slightly from the previously reported values for basins with low or high mean elevation when $a$ is estimated, but varies somewhat in the transition zone where $a$ has a larger effect. While the fall in ERL is appreciable in the transition zone, the ERL remain high, ranging from 98 to 86 years depending on duration, which are associated with variance of prediction ranging from 0.077 to 0.086. Thus, the choice of NL scale parameter does influence the variance of prediction for basins in the narrow transition zone between high and low elevation basins, but does not seriously impact regional skew model form or the fitted regression coefficients. The extended model may do better in some cases because of a smaller

model error variance, $\sigma_\delta^2$, but generally has a larger MSE because of the larger parameter area that is now included.

**Table 4.8:** Comparison of Nominal Effective Record Length (ERL) for the final non-linear elevation regional skew model (Equation (4.10)) and the extended model (Equation (4.12)).

| | 1-Day | | 3-Day | | 7-Day | | 15-Day | | 30-Day | |
|---|---|---|---|---|---|---|---|---|---|---|
| Elevation | Final Model | Extended Model | Final Model | Extended Model | Final Model | Extended Model | Final Model | Extended Model | Final Model | Extended Model |
| < 2500 | 186 | 192 | 172 | 174 | 156 | 157 | 157 | 154 | 145 | 148 |
| 3000 | 182 | 180 | 168 | 163 | 155 | 146 | 156 | 141 | 144 | 137 |
| 3200 | 177 | 157 | 164 | 142 | 153 | 129 | 155 | 123 | 144 | 121 |
| 3400 | 170 | 121 | 159 | 112 | 151 | 104 | 154 | 100 | 143 | 99 |
| 3600 | 164 | 98 | 155 | 93 | 151 | 91 | 154 | 88 | 142 | 86 |
| 3800 | 162 | 113 | 155 | 109 | 153 | 108 | 156 | 104 | 141 | 101 |
| 4000 | 162 | 158 | 157 | 148 | 156 | 145 | 157 | 139 | 141 | 133 |
| >4500 | 162 | 169 | 157 | 159 | 156 | 156 | 157 | 150 | 140 | 142 |

Figure 4.20 plots the ratio of the ERL of the Extended model and the ERL of the Final model (i.e. ERL(Ext. Model)/ERL(Final Model)) across a range of basin elevations. For each duration, the ratio is nearly equal to one for low and high elevations, indicating that the ERL of the extended model and the ERL of the Final model are nearly equal. In the transition zone between 3,000 ft and 4,000 ft, the ratio is much less than one, indicating that the ERL for the extended model is less than the ERL for the Final model.

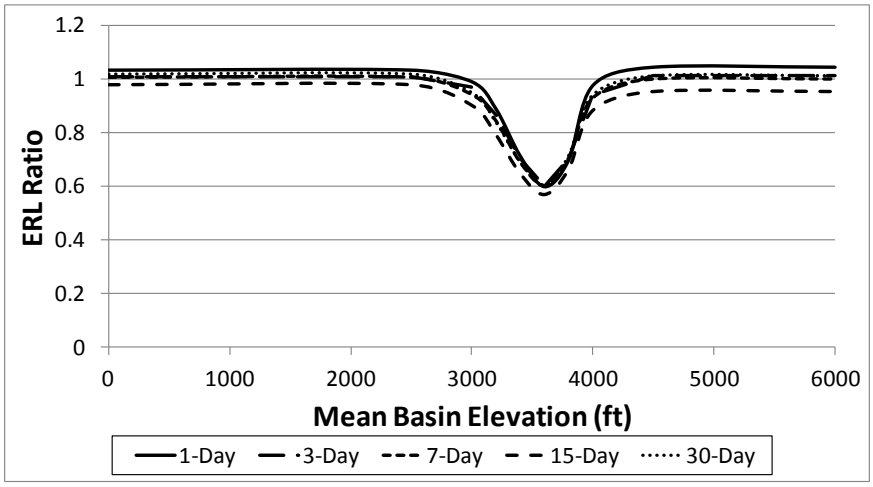


**Figure 4.20:** Ratio of ERL from the Extended model and ERL from the final model (ERL(Ext. Model)/ERL(Final Model)) versus mean basin elevation for five study durations.

This analysis confirmed that using a single non-linear parameter *a* for consistency across durations was statistically justified because the first-order linearization term was not significant at any duration. Moreover, except in the immediate zone of the transition, treating the universal $a$ as a known quantity would have very little effect in the computed ERL of skew estimates.

***Conclusion***

The EMA was used to estimate at-site sample log-space skew coefficients for 1-day, 3-day, 7-day, 15-day, and 30-day duration rainfall floods for 50-sites in and around the central valley of California. Bayesian Generalized Least Squares regression failed to provide stable regional skew coefficient models due to extremely high cross-correlation between sampling error of the skew coefficients. A Bayesian WLS/GLS analysis was developed and implemented which utilized weighted least squares to develop model parameters and generalized least squares to estimate their precision. Strong non-linear trends in mean basin elevation were observed, and a non-

linear elevation term similar to that used in the previous California instantaneous annual maximum study was utilized.

Regression results were very good, with ERL ranging from 140 to 192 years depending on rainfall flood duration and mean basin elevation. Leverage and influence statistics were calculated for each regional model. No basin had very high leverage or influence at all durations. A linearization of the non-linear elevation term confirmed the validity of using a common non-linear scale parameter $a$ across all durations, rather than a unique $a$ parameter for each duration.

# REFERENCES


Cohn, T.A., W.L. Lane, J.R. Stedinger, 2001, Confidence intervals for Expected Moments Algorithm flood quantile estimates: Water Resources Research, v. 37, no. 6, 1695-1706, doi: 2001WR900016.

Draper, N.R. and Smith, H. (1967). *Applied Regression Analysis*. John Wiley & Sons, Inc., New York, N.Y.

Feaster, T.D., Gotvald, A.J., and Weaver, J.C., 2009, Magnitude and frequency of rural floods in the southeastern United States, 2006—Volume 3, South Carolina: U.S. Geological Survey Scientific Investigations Report 2009-5156, 226 p.

Griffis, V. W. 2003. "Evaluation of Log-Pearson type 3 flood frequency analysis methods addressing regional skew and low outliers." MS thesis, School of Civil and Environmental Engineering, Cornell Univ., Ithaca, N.Y.

Griffis, V. W. , J.R. Stedinger, and T. A. Cohn . (2004). "LP3 Quantile Estimators with Regional Skew Information and Low Outlier Adjustments", *Water Resources Research, 40*, W07503, doi:1029/2003WR002697.

Griffis, V.W., and J. R. Stedinger, (2009), The Log-Pearson Type 3 Distribution and its Application in Flood Frequency Analysis, 3. Sample Skew and Weighted Skew Estimators, *J. of Hydrol. Engineering* 14(2), pp. 121-130.

Gruber, A.M. and J.R. Stedinger, (2008), Models of LP3 Regional Skew, Data Selection and Bayesian GLS Regression, Paper 596, World Environmental and Water Resources Congress – Ahupua'a, Babcock, R.W. and R. Watson editors, Honolulu, Hawai'i, May 12-16.

Gotvald, A.J., Feaster, T.D., and Weaver, J.C., 2009, Magnitude and frequency of rural floods in the southeastern United States, 2006—Volume 1, Georgia: U.S. Geological Survey Scientific Investigations Report 2009–5043, 120 p.

Interagency Advisory Committee on Water Data, 1982, Guidelines for determining flood-flow frequency, Bulletin #17B of the Hydrology Subcommittee, Office of Water Data Coordination: U.S. Geological Survey, Reston Virginia, 183 p. Available at *http://water.usgs.gov/osw/bulletin17b/dl_flow.pdf*

Lamontagne, J.R., Stedinger, J.R., Berenbrock, Charles, Veilleux, A.G., Ferris, J.C., and Knifong, D.L., 2012, Development of regional skews for selected flood durations for the Central Valley Region, California, based on data through water year 2008: U.S. Geological Survey Scientific Investigations Report 2012–5130, 60 p.

Martins, E.S., and Stedinger, J. R., 2002, Cross-correlation among estimators of shape: Water Resources Research, v. 38, no. 11, 1252, doi: 10.1029/2002WR001589

Parrett, C., A. Vellieux, , J. R. Stedinger, N. A. Barth, D. Knifong, , and J.C. Ferris, 2010. Regional Skew for California and Flood Frequency for Selected Sites in the Sacramento-San Joaquin River Basin Based on Data through Water Year 2006, OFR 2010-5260, U.S. Geological Survey.

Reis, D.S., Jr., Stedinger, J.R., and Martins, E.S., 2005, Bayesian generalized least squares regression with application to the log Pearson type III regional skew estimation: Water Resources Research, 41, W10419, doi:10.1029/2004WR003445.

Stedinger, J. R., 1983, Estimating a Regional Flood Frequency Distribution: Water Resources Research, v. 19, no. 2, p. 503-510.

Stedinger, J. R., and Tasker, G. D., 1985, Regional hydrologic analysis, 1, ordinary, weighted and generalized least squares compared: Water Resources Research ,v. 21, no. 9, p. 1421-1432. [with correction, Water Resources Research, v. 22, no. 5, p. 844, 1986.]

Tasker, G.D., and J.R. Stedinger, 1989. An Operational GLS Model for Hydrologic Regression, Journal of Hydrology, 111(1-4), 361–375.

Weaver, J.C., Feaster, T.D., and Gotvald, A.J., (2009), Magnitude and frequency of rural floods in the southeastern United States, 2006—Volume 2, North Carolina: U.S. Geological Survey Scientific Investigations Report 2009–5158, 113 p.

Veilleux, A. G. 2009. "Bayesian GLS Regression for Regionalization of Hydrologic Statistics, Floods and Bulletin 17 Skew." MS thesis, School of Civil and Environmental Engineering, Cornell Univ., Ithaca, N.Y.

# CHAPTER 5

# MODEL PRECISION, ANALYSIS OF VARIANCE, AND MODEL CONSISTENCY

This chapter explores several concerns which were raised during the USGS/USACE review process for Lamontagne et al. [2012] (Chapter 4 includes those results). One concern was that the effective record lengths reported in Chapter 4 were much greater than the previous California annual peak skew study [Parrett et al., 2011]. A related concern is that the reported variance of predictions were very small. These concerns are addressed here in Section 5.1. A second concern was that the computed sums of squared deviations from the mean were much smaller than the estimate in the Pseudo ANOVA table for every duration. This concern is explored here in Section 5.2. That discussion raises the general issue of how an ANOVA or a pseudo ANOVA should be constructed to explain how variability in the data can be partitioned among variability explained by the model, true unexplained variability, and sampling error. Section 5.3 addresses the concern that the skew models do not seem to trend in duration, which might lead to inconsistencies in the subsequent flood frequency analysis.

## *Section 5.1: Effective Record Length and Regional Skew*

The results in Chapter 4 pertaining to the regional skew analysis for California rainfall floods are published in Lamontagne et al. [2012]. When that report was in review, some surprise and concern was raised over the remarkably low variance of prediction ($VP_{new}$) and high effective record length (ERL). This section explores

whether the results reported in Chapter 4 are reasonable and seeks to answer the concerns which were raised. This is done by first considering how ERL is computed, and what factors can influence its value. Second, a very simple approximation of the final non-linear models in Chapter 4 is constructed, and a simple approximation to its $VP_{new}$ is compared to the reported $VP_{new}$ from Chapter 4. Finally, the discussion highlights the differences between the analysis in Chapter 4 and previous GLS studies: in particular the discussion considers the previous California instantaneous annual maximum skew study [Parrett et al., 2011].

***Section 5.1.1: Computation of the Effective Record Length***

The Effective Record Length (ERL) is an index by which to understand the implications of the magnitude of $VP_{new}$. If the ERL of the regional skew model is greater than the at-site record length, then regional information will be more important than the at-site data, and vice versa.

With the Bulletin 17B frequency procedure, the variance of prediction is used to assign weights to the regional skew and the at-site skew reflecting the relative precision of each [IACWD, 1982]. The ERL values in Table 4.5 are computed using the Griffis-Stedinger formula for the mean square error of the skew coefficient [Griffis and Stedinger, 2009], their formula can be written:

$$MSE[\hat{\gamma}] = \left(\frac{6}{N} + a(N)\right)\left(1 + \left\{\frac{9}{6} + b(N)\right\}\gamma^2 + \left\{\frac{15}{6*8} + c(N)\right\}\gamma^4\right)$$

where $\hat{\gamma}$ is the sample skew, $\gamma$ is the true skew, and a, b, and c are correction factors for small sample sizes. ERL is computed by setting $\gamma$ equal to the regional skew

coefficient, $\gamma_R$, and solving for the $N$ such that $MSE[\hat{\gamma}|\gamma_R] = VP_{new}$. Griffis and Stedinger [2009] provide an alternative ERL statistic based on the ratio of the MSE of the regional and the at-site skew coefficient.

Two factors enter the ERL computation: the regional skew coefficient value, $\gamma_R$, and the variance of prediction for a site, $VP_{new}$. Figure 5.1 plots the Griffis-Stedinger MSE of the sample skew coefficient, $\hat{\gamma}$, versus the regional skew coefficient, $\gamma_R$, for three record lengths. Here $\gamma_R$ is serving as the unknown population skew. The strong dependence of the MSE on the assumed $\gamma_R$ is evident. Note that two sample skew coefficients with the same record length, but different $\gamma_R$ skews, have different MSE.

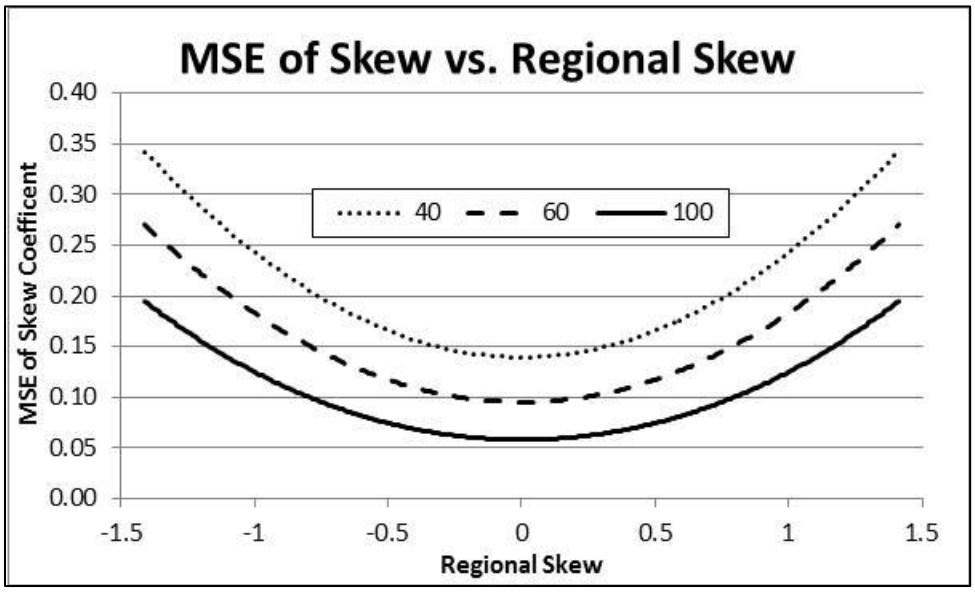


**Figure 5.1:** MSE of the Skew Coefficient versus Regional Skew used to compute MSE

Again using the relationship in Griffis and Stedinger [2009], Figure 5.2 plots the ERL versus MSE for $\gamma_R$. For a constant MSE, as $\gamma_R$approaches zero, the ERL

approaches the minimum for that MSE. This follows from Figure 5.2 wherein the $\gamma_R = 0$ line has the smallest ERL for any fixed MSE.

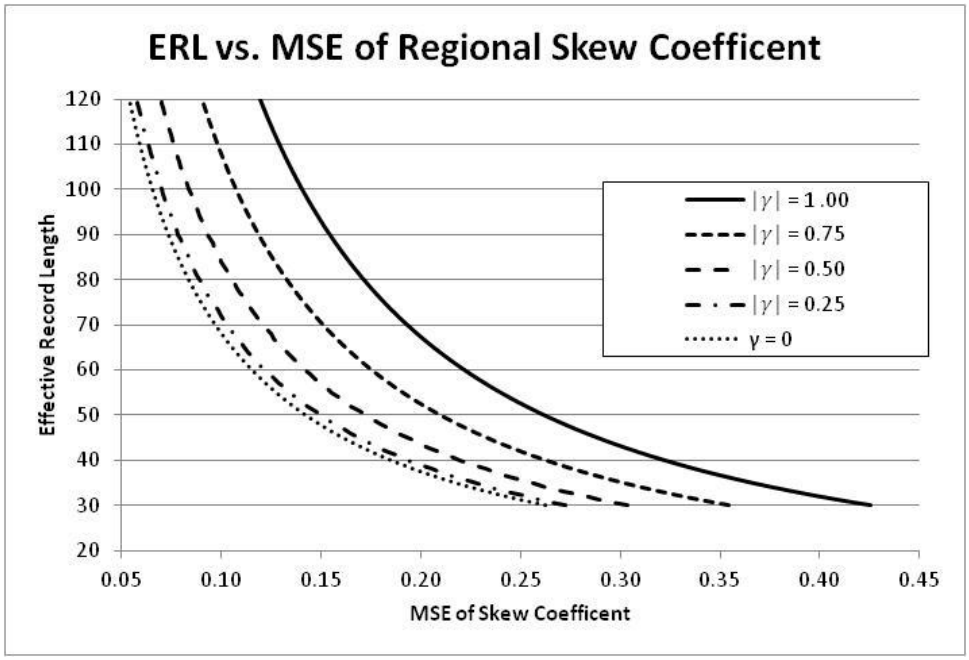


**Figure 5.2:** ERL versus MSE of Regional Skew Coefficient

This effect revealed itself in other regional skew studies. Consider the Southeast Skew Study [Veilleux, 2009] and the California Annual Maximum Skew Study [Parrett et. al., 2011]. Both studies reported an average variance of prediction of 0.14, but the Southeast Skew Study reported an ERL of 40 years, whereas the California Annual Maximum Skew Study reported an ERL of 60 years. This difference is because the regional model in the Southeast is a constant of -0.019, while the California regional skew model takes values from [-0.62, 0.68]. Similarly, the regional skew analysis for Iowa resulted in a constant model equal to -0.4, with an average variance of prediction of 0.16, and an ERL of 50 years [Eash, 2013]. Thus, with a larger average variance of prediction, the Iowa study reports a larger ERL than the Southeast study with a smaller average variance of prediction.

Thus it is not surprising that ERL values reported in Chapter 4 are greater than those reported in previous studies: the magnitude of the regional skew was greater than that reported in the Southeast for all durations and also greater than the California annual maximum model for most durations. This also explains in part why the ERL decreases with increasing duration and increasing elevation: this occurs because the regional skew magnitude is greatest at short durations and low elevations and the same MSE for the assumed skew will have a larger ERL because a larger sample would be required to achieve the MSE value.

***Section 5.1.2: Variance of Prediction for a Simple Regional Skew Model***

The previous section shows that the high ERL reported in Chapter 4 are in part due to the greater magnitude of the California duration skew model relative to previous studies. Still, the reported variances of prediction in this study are much smaller than in Parrett et al. [2011]. As a check of the WLS/GLS analysis in Chapter 4, consider the following simple analysis of the data.

Given the steep rise between the low and high elevation sites, the final models in Chapter 4 are essentially a constant average skew for low elevation sites and a second constant average skew for high elevation sites. In the WLS/GLS regression in Chapter 4, these averages are weighted means where the weights are essentially based on record length for each of the sites, because the model error is so small and cross-correlation of sampling errors is ignored.

For the purposes of the following discussion, define three elevation classes: L) low (mean basin elevation < 3,500 ft), H) high (mean basin elevation > 4,000 ft), and

M) medium (mean basin elevation [3,500 ft, 4,000 ft]. This results in $n_L$= 24 class L sites, $n_H$= 20 class H sites, and $n_M$= 6 class M sites. In the simple analysis that follows, the medium elevation basins (class M) are neglected.

A simple approximation of the final model reported in Chapter 4 can be based upon:

$$\hat{y}_j = \frac{1}{\sum_{i \in I_j} W_j(i)} \sum_{i \in I_j} W_j(i) \hat{\gamma}_i \qquad (5.1)$$

where,

$j$ is an elevation class taking either $j = L$ for low elevation or $j = H$ for high elevation sites,
$\hat{y}_j$ is the regional skew coefficient for elevation class $j = L, H$
$n_j$ is the number of sites in elevation class $j = L, H$
$I_j$ is a set of site indices for elevation class $j = L, H$
$\hat{\gamma}_i$ is the sample skew coefficient for site $i$, and
$W_j(i)$ is the weight for site $i$ which is a member of elevation class $j = L, H$.

Let $\sum(j)$ be the $n_j \times n_j$ covariance matrix of the $n_j$ sample skews in elevation class $j$.

Let $\Sigma_{WLS}(j)$ be a $n_j \times n_j$ matrix containing only the diagonal elements of $\sum(j)$.

The weight vector $W_j$ employed in this simple analysis using ( 5.1) has the value:

$$W_j = \left(e_j{}^T \Sigma_{WLS}(j)^{-1} e_j\right)^{-1} e_j{}^T \Sigma_{WLS}(j)^{-1}, \; j = L \; or \; H \qquad (5.2)$$

where $e_j{}^T$ = (1 ... 1) is a $1 \times n_j$ row vector of ones. The $i^{th}$ entry of $W_j$ is the weight for site $i$ in elevation class $j$, $W_j(i)$. This is a special and simple case of the standard WLS weight matrix for a constant model, ignoring model error variance. If one considers the case that the observed sample skew coefficients are correlated, the sampling variance of the constant value for model, $j = L$ or $H$, described by ( 5.1) is:

$$V_j = \frac{W_j \Sigma_j W_j^T}{\left(\sum w_j(i)\right)^2} \quad (5.3)$$

For any model, variance of prediction is composed of two elements: model error variance and sampling error variance of the fitted model [Reis et al., 2005]. An approximation of the variance of prediction for this simple analysis is given by:

$$\widetilde{VP}(j) = V_j + \sigma_\delta^2 \quad (5.4)$$

where $\sigma_\delta^2$ is the model error variance assigned to the final non-linear elevation model by the Bayesian GLS procedure described in Chapter 4. Let $VP_{new}(i)$ be the variance of prediction for the final model in the Bayesian GLS analysis in Chapter 4 at site $i$. If $\widetilde{VP}(j)$ is nearly equal to $VP_{new}(i)$ for $i \in I_j$ it indicates that, given the small model error variance, the reported $VP_{new}$ and ERL from the complicated analysis in Chapter 4 are consistent with the much simpler analysis reported here. This is desirable because it would indicate the 'back-of-the-envelope' analysis described above confirms the complex Bayesian GLS analysis in Chapter 4, given the computed model error variance.

The preceding analysis is consistent with the procedure in Chapter 4 because the model was estimated through a WLS analysis and the precision of that model was assessed from a GLS analysis, i.e. the cross-correlation of the sampling errors was neglected in the model selection, but was considered when assessing its precision. A potential drawback of the simple analysis is that it ignores the contribution of the model error variance to the sampling error variance of the model parameter. For

simplicity, it is assumed to be zero in ( 5.1)-( 5.3) and in the definition of $\Sigma_j$. However the model error variance is much smaller than the average sampling error variance (see ASVE in Table 4.3), so this is not expected to have a significant impact on the magnitude of $V_j$.

The value of $VP_{new}(i)$ for the final models in Chapter 4 varies from site to site. For this reason, their average across all study sites, $AVP_{new}$, is reported in Table 4.3 in Chapter 4. The simple analysis in this section divides sites into elevation classes, and computes a unique $\hat{y}_j$ and $\widetilde{VP}(j)$ for $j = L$ and $H$. Thus consider an elevation class specific $AVP_{new}$ based on $VP_{new}(i)$ values computed with the Chapter 4 model:

$$AVP_{new}(j) = \frac{1}{n_j}\sum_{i \in I_j} VP_{new}(i) \qquad (5.5)$$

Here $VP_{new}(i)$ is the $VP_{new}$ for site $i$ from the analysis in Chapter 4. It is expected that $AVP_{new}(j)$ will be nearly equal to $\widetilde{VP}(j)$. This will help show that the small $AVP_{new}$ reported in Chapter 4 are consistent with the $\widetilde{VP}(j)$ for the very simple, straightforward analysis in this section. Table 5.1 reports the values of $V_j$ and $\widetilde{VP}(j)$ for the simple model, and $AVP_{new}(j)$ from ( 5.5) for each duration and for low ($j = L$) and high ($j = H$) elevation classes. Figure 5.3 and Figure 5.4 plot the data in Table 5.1 for the low and high elevation classes respectively.

**Table 5.1:** Comparison of $\boldsymbol{V_j}$ and $\boldsymbol{\widetilde{VP}(j)}$ from simple weighted mean analysis and reported $\boldsymbol{AVP_{new}(j)}$ from Chapter 4 for low and high elevation categories and for five study durations.

| Duration | 1-Day | | 3-Day | | 7-Day | | 15-Day | | 30-Day | |
|---|---|---|---|---|---|---|---|---|---|---|
| Elevation ($j$) | Low ($j$=1) | High ($j$=2) | Low ($j$=1) | High ($j$=2) | Low ($j$=1) | High ($j$=2) | Low ($j$=1) | High ($j$=2) | Low ($j$=1) | High ($j$=2) |
| $V_j$ | 0.045 | 0.028 | 0.048 | 0.030 | 0.049 | 0.033 | 0.054 | 0.035 | 0.054 | 0.037 |
| $\widetilde{VP}(j)$ | 0.056 | 0.038 | 0.057 | 0.039 | 0.056 | 0.040 | 0.059 | 0.040 | 0.063 | 0.046 |
| $AVP_{new}(j)$ | 0.056 | 0.039 | 0.057 | 0.040 | 0.057 | 0.041 | 0.060 | 0.041 | 0.064 | 0.048 |

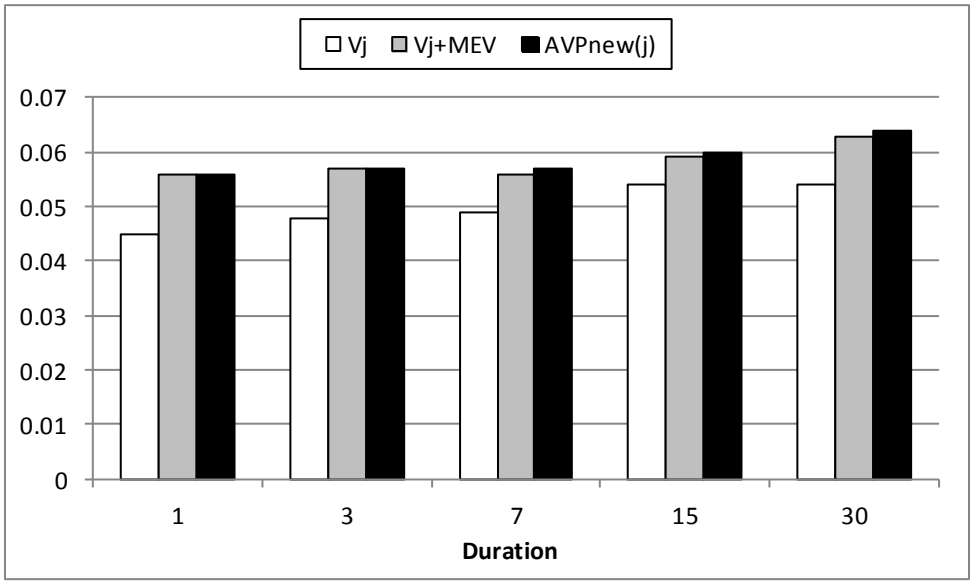


**Figure 5.3:** Low elevation $\boldsymbol{V_j}$, $\boldsymbol{\widetilde{VP}(j)}$, and $\boldsymbol{AVP_{new}(j)}$ for five study durations.

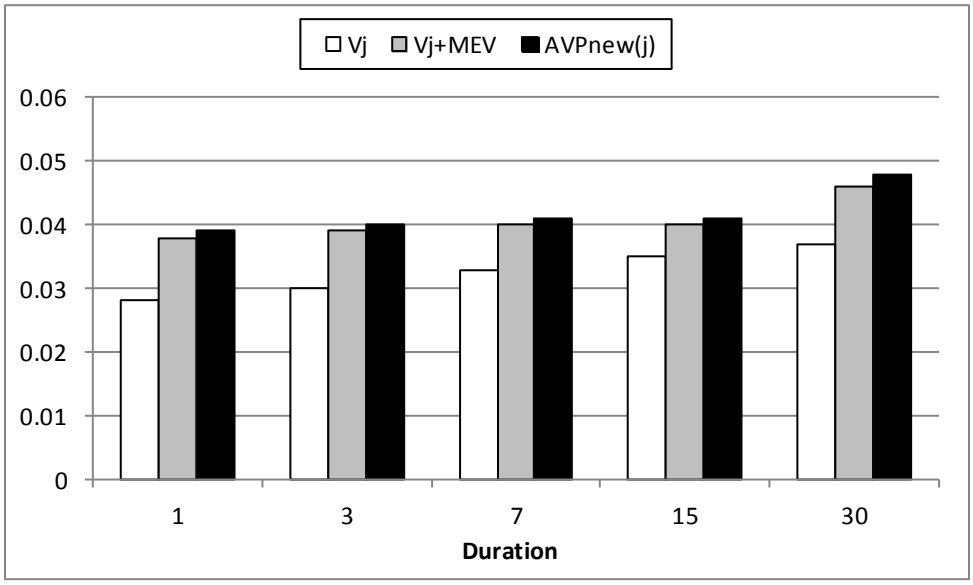


**Figure 5.4:** High elevation $\boldsymbol{V_j}$, $\boldsymbol{\widetilde{VP}(j)}$, and $\boldsymbol{AVP_{new}(j)}$ for five study durations.

As was expected, the approximation of the variance of prediction described in ( 5.4) are nearly equal to $AVP_{new}(j)$ in all durations and at both high and low elevations. In a few cases $\widetilde{VP}(j) < AVP_{new}(j)$. This is likely because $V_j$ does not consider the contribution of model error variance to the sampling error of the computed model. As expected, these differences are very small. Importantly, Table 5.1 indicates that the remarkably low $AVP_{new}$ reported by the GLS analysis in Chapter 4 are very reasonable, given the reported small model error variances, resulting in the dominance of the sampling errors. Even modest changes in the model error variance would not change the $AVP_{new}(j)$ as long as the sampling error continues to dominate.

### *Section 5.1.3: Comparison of Results from Chapter 4 and CA Annual Maximum Study*

The analysis in the previous section was initially intended to assure reviewers for Lamontagne et al. [2012] that the low $VP_{new}$ and high ERL associated with the final models in that report are reasonable. Many of their concerns centered on

comparisons between Lamontagne et al. [2012] and the previous California Annual Maximum Skew Study (Parrett et al. [2011]). Parrett et al. found an $AVP_{new}$ of 0.14 with a nominal ERL of 60 years versus the $AVP_{new}$ of 0.048-0.056 and ERL of 133-150 in Chapter 4, even though the geographic area of the two studies was similar.

A significant difference between the two studies was that they modeled different phenomena. Parrett et al. [2011] considered instantaneous peak flows of all sources. Lamontagne et al. [2012] considered duration flows caused primarily by rainfall.

Even the 1-Day annual maximum rainfall flood series and the instantaneous annual maximum series are very different. First, all snowmelt floods have been removed from the flow record. Second, the rainfall flood series has been averaged over a 24-hour period. The two phenomena will have different distributions, potentially very different distributions, and consequently different skew coefficients. This difference will likely increase with duration, i.e. the 1-Day rainfall flood record is likely more similar to the instantaneous peaks than the 30-Day rainfall flood record. Thus comparisons between the Parrett et al. [2011] and Lamontagne et al. [2012] studies must be made cautiously.

Another significant difference between the two studies is the number and type of sites included. The majority of the 50 sites included in Lamontagne et al. [2012] were large basins associated with US Army Corps of Engineers dams, while the 158 sites in Parrett et al. [2011] run the gamut from small mountain streams to larger rivers. Thus, there was much more real variation to describe in the Parrett et al. [2011]

study. This is demonstrated by comparing the model error variance of the constant model, $\sigma_\delta^2(0)$, from the two studies. $\sigma_\delta^2(0)$ describes the expected variation in the true skew coefficients. Parrett et al. [2011] report a $\sigma_\delta^2(0)$=0.2, while Lamontagne et al. [2012] report $\sigma_\delta^2(0)$ = [0.03,0.08] depending on duration. There can be almost an order of magnitude less expected variation in the true skews in the Lamontagne et al. [2012] analysis than the Parrett et al. [2011] analysis.

Thus it is not surprising that the MEV for the final model, $\sigma_\delta^2(k)$, in Lamontagne et al. [2012] is much smaller than in Parrett et al. [2011] (about 0.01 for all durations versus 0.1). Recall from Chapter 3 (equation (3.5)) that $VP_{new}$ can be represented as:

$$VP_{new} = Expected\ Model\ Error\ Variance + Expected\ Sampling\ Error\ Variance$$

The average sampling error variance for the final models in Parrett et al. [2011] and Lamontagne et al. [2012] are nearly equal (0.03 vs. [0.04-0.05] depending on duration). The large difference in the $VP_{new}$ from the two studies is mostly attributable to the difference in the MEV, which seems to be largely driven by the nature of the phenomena modeled and the types of sites included.

***Conclusion:***

Section 5.1 addressed concerns raised during the review of Lamontagne et al. [2012] (which is largely reproduced in Chapter 4) that the reported $VP_{new}$ and ERL are unrealistic, particularly when compared to previous skew studies. These concerns

are addressed in three ways. Section 5.1.1 discusses the effect of the regional skew magnitude on computed ERLs. The magnitude of the regional skew values reported in this study are generally greater than in previous studies.

Section 5.1.2 presents a very simple analysis as an approximation of the more complicated analysis presented in Chapter 4. The approximate variance of prediction from this analysis, $\widetilde{VP}(j)$, is compared to average values of the $VP_{new}$ values reported in Chapter 4. It was found that the simple analysis agrees well with the more complicated analysis, reassuring us that $VP_{new}$ is reasonable. Finally, Section 5.1.3 discusses the previous California skew study and the analysis in Chapter 4. It is observed that: 1) they are modeling very different phenomena making comparisons difficult at best, and 2) the range of study basins considered in Chapter 4 limits the amount of variability in the true skew coefficients and thus the magnitude of the model error variance is very small, resulting in small $VP_{new}$ and large ERL.

***Section 5.2: Analysis of Variance re-examined***

ANalysis Of Variance (ANOVA) and the coefficient of determination are used in regression studies to quantify the overall variability in a dataset and the ability of models to explain that variability. Because of the unique characteristics of the regional GLS studies reported in Reis et al., [2005], Chapter 4, and several USGS reports, a pseudo $R^2$ and a pseudo ANOVA was developed. The generated statistics did not include in the ANOVA table a correction for cross-correlation between estimators of the hydrologic statistics of interest, and that could be a concern. Moreover, in the regional skew study described in Chapter 4, results of the pseudo ANOVA (Table 4.4) varied widely from those of a traditional ANOVA. In some

cases the discrepancy in the estimated total sum of squared deviation was as much as 40%. These concerns led to a general re-examination of the ANOVA issue, which is described in the following section.

Various ANOVA have been proposed for generalized least squares regression (see Greene [2008] for a summary). The following discussion will divide proposed GLS ANOVA into empirical and theoretical approaches. Empirical approaches estimate variability directly from the data. The traditional ANOVA for ordinary least squares is thus an empirical approach. Theoretical approaches estimate the division of variability using the theoretical or assumed error structure for the data and estimated parameters, particularly various variance components. The pseudo ANOVA recommended by Gruber et al. [2007] and employed in Chapter 4 is an example of a theoretical approach.

A third ANOVA approach arises implicitly from a family of "pseudo $R^2$" statistics based upon maximum likelihood analysis. These MLE-based $R^2$ statistics are computed as the ratio (or some transformation of the ratio) of the likelihood function values of the final model and the constant model [see for example Baxter and Cox, 1970]. Such statistics are commonly applied in logistic regression [Menard, 2000] and discrete-count regression models including Poisson and negative-binomial regression [Dobson, 2002; Cameron and Trivedi, 1998; Liu et al., 2005]. A pseudo ANOVA table can generally be constructed to match computed values of a pseudo $R^2$ statistic. Such MLE-based pseudo-$R^2$ tables are not explored in detail in this section.

Before examining the empirical and theoretical approaches, it is important to understand the difference in their conceptual motivation. The empirical ANOVA

divides the observed variation for the sample among several components corresponding to variation explained by the model, and residual unexplained errors. The theoretical approach can be more general: can describe the expected variation in a random sample with the estimated variance and correlation structure. Thus, the theoretical approach is somewhat divorced from the actual sample, and its results deviate from the empirical approach. The reasons for the development and use of theoretical approaches over empirical ones are explored in this section.

As a final introductory aside, it is important to note how the various ANOVA in this section relate to ANOVA types I, II, and III [Herr, 1986]. Herr describes how ANOVA types I, II, and III can be used to describe the effect of different treatments in an experiment. For example, suppose a model of body mass index (BMI) was based on two treatments: whether a person exercised and whether a person ate a healthy diet. ANOVA types I, II, and III each seek to explain the incremental amount of variability explained by each individual treatment. The difference between the three "types" lies the underlying assumptions about treatment hierarchy and interaction. See Herr [1986] for an in-depth discussion. Type I assumes a hierarchy to the treatments, and adds each treatment to the model in order of importance. The incremental sum of squares explained by each treatment is reported, as well as the sum of squares for its interaction with the previously added treatments. Type II ANOVA does not assume a hierarchy, but reports the incremental sum of squares explained by each treatment given all other treatments are already included in the model. Type II ANOVA assumes no interactions between the treatments. Type III ANOVA reports the incremental sum of squares explained by each treatment, assuming all other treatments and their

interactions are already included in the model. These ANOVA methods date to the 1930s, though the "type I, II, and III" terminology is traced to the SAS software [Langsrud, 2003]. In contrast to those methods, each of the different ANOVA described here are concerned with the fraction of variability explained by the whole model, and not with the incremental effect of each explanatory variable. The different ANOVA presented in this section address the division of variability between models and errors, and do not address assumptions about the model structure or the hierarchy of explanatory variables.

Section 5.2 has several subsections. Section 5.2.1 describes empirical ANOVA approaches which have been applied in both ordinary and generalized least squares regression studies. Section 5.2.2 describes the development of theoretical ANOVA approaches for GLS by Reis [2005] and Gruber et al. [2007] for application with regional hydrologic studies. Section 5.2.3 considers a new theoretical ANOVA approach for such hydrologic studies. Section 5.2.4 provides a numerical example describing a hydrologic study and discussion comparing two theoretical ANOVA approaches with an empirical ANOVA approach. Finally, Section 5.2.5 concludes with a discussion of the relative merits of the various approaches for describing the partition of variability among that explained by a model, that due to model error, and that due to sampling error in regional hydrologic studies.

***Section 5.2.1: Empirical ANOVA Approaches***

The analysis in Chapter 4 focused on regional skew regression. For greater generality, this section will frame the discussion in terms of a dependent variable $y_i$. The observed value of $y_i$ is $\tilde{y}_i$. This distinction is needed because in many

applications, such as regional skew regression, only estimates of $y_i$ are available. The sample mean of a given set of $\tilde{y}_i$ is $\bar{y}$. The least squares model estimate of $y_i$ is $\hat{y}_i$. The following discussion will assume a linear model of the form:

$$\tilde{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{5.6}$$

Where

$\tilde{\boldsymbol{Y}}$ is a $n \times 1$ vector of $\tilde{y}_i$'s, $\boldsymbol{X}$ is a $n \times k$ matrix of independent variables,

$\boldsymbol{\beta}$ is a $k \times 1$ vector of decision variables, and
$\boldsymbol{\varepsilon}$ is a $n \times 1$vector of errors.

It is assumed that the errors have zero mean, so that $E[\boldsymbol{\varepsilon}] = 0$, and covariance matrix $E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T] = \boldsymbol{\Lambda}$. The error covariance matrix can take three forms depending on the type of regression employed.

In ordinary least squares (OLS) regression, the errors are assumed to be uncorrelated and homoscedastic, thus

$$\boldsymbol{\Lambda}_{\text{OLS}} = \mathbf{I}\sigma^2 \tag{5.7}$$

where $\sigma^2$ is the error variance and $\mathbf{I}$ is a $n \times n$ identity matrix.

In a weighted least squares regression (WLS), the errors are assumed to be uncorrelated and heteroscedastic, thus

$$\begin{aligned} \boldsymbol{\Lambda}_{\text{WLS}}(i,k) &= \sigma_i^2 \quad \text{if } i = k \\ \boldsymbol{\Lambda}_{\text{WLS}}(i,k) &= 0 \quad \text{if } i \neq k \end{aligned} \tag{5.8}$$

where $\sigma_i^2$ is the variance of $\varepsilon_i$.

Finally, In generalized least squares (GLS), the errors are assumed to be correlated and heteroscedastic, thus

$$\begin{aligned} \boldsymbol{\Lambda}_{\mathrm{GLS}}(i,k) &= \sigma_i^2 \qquad \text{if } i = k \\ \boldsymbol{\Lambda}_{\mathrm{GLS}}(i,k) &= \rho_{i,k}\sigma_i\sigma_k \quad \text{if } i \neq k \end{aligned} \tag{5.9}$$

where $\rho_{i,k}$ is the cross correlation of $\varepsilon_i$ and $\varepsilon_k$. Note that this definition will be adjusted slightly in the subsequent analysis of the Stedinger and Tasker [1985] hydrologic regional regression framework. In that framework the $\boldsymbol{\Lambda}$ is composed of sampling errors which are correlated, and a constant model error variance which is not correlated.

Recall from Chapter 3 that the least squares estimate of the model parameters $\boldsymbol{\beta}$ is given by

$$\boldsymbol{b}_J = \left(\boldsymbol{X}^T\boldsymbol{\Lambda}_J^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{\Lambda}_J^{-1}\widetilde{\boldsymbol{Y}} \tag{5.10}$$

where $J$ is the regression case (either OLS, WLS, or GLS). Let $\boldsymbol{W}_J$ be the weight matrix for the $J$ case, defined as:

$$\boldsymbol{W}_J = \left(\boldsymbol{X}^T\boldsymbol{\Lambda}_J^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{\Lambda}_J^{-1} \tag{5.11}$$

Thus, the β-parameter estimator can be written $\boldsymbol{b}_J = \boldsymbol{W}_J\widetilde{\boldsymbol{Y}}$.

***Section 5.2.1.1 Untransformed Empirical ANOVA***

The standard ANOVA table reports the total sum of squared errors about the mean, $SS_T$ as a measure of the overall variability in the data. $SS_T$ is computed as:

$$\begin{aligned} SS_T &= \sum_{i=1}^{n}(\tilde{y}_i - \bar{y})^2 = \left(\widetilde{\boldsymbol{Y}} - \bar{y}\boldsymbol{e}\right)^T\left(\widetilde{\boldsymbol{Y}} - \bar{y}\boldsymbol{e}\right) \\ &= \left(\widetilde{\boldsymbol{Y}} - \overline{\boldsymbol{Y}}\right)^T\left(\widetilde{\boldsymbol{Y}} - \overline{\boldsymbol{Y}}\right) \end{aligned} \tag{5.12}$$

where

$\bar{y}$ is the sample mean of $\widetilde{\boldsymbol{Y}}$,
$\boldsymbol{e}$ is a $n \times 1$ vector of ones, and
$\overline{\boldsymbol{Y}}$ is a $n \times 1$ vector containing $\bar{y}$ for every element.

Let $\widehat{\boldsymbol{Y}}_J$ be a $n \times 1$ vector containing the fitted model estimates of $\widetilde{\boldsymbol{Y}}$ for regression case $J$. Expanding (5.12):

$$\begin{aligned} SS_T &= \left(\widetilde{\boldsymbol{Y}} - \overline{\boldsymbol{Y}}\right)^T \left(\widetilde{\boldsymbol{Y}} - \overline{\boldsymbol{Y}}\right) \\ &= \left(\widetilde{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_J + \widehat{\boldsymbol{Y}}_J - \overline{\boldsymbol{Y}}\right)^T \left(\widetilde{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_J + \widehat{\boldsymbol{Y}}_J - \overline{\boldsymbol{Y}}\right) \\ &= \left(\widetilde{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_J\right)^T \left(\widetilde{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_J\right) + \left(\widehat{\boldsymbol{Y}}_J - \overline{\boldsymbol{Y}}\right)^T \left(\widehat{\boldsymbol{Y}}_J - \overline{\boldsymbol{Y}}\right) + 2\left(\widetilde{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_J\right)^T \widehat{\boldsymbol{Y}}_J - \overline{\boldsymbol{Y}} \\ SS_T &= SS_E(J) + SS_M(J) + 2CP(J) \end{aligned} \tag{5.13}$$

Thus, the $SS_T$ is divided into three components: *error* sum of squares $SS_E$, *model* sum of squares $SS_M$, and a *cross-product* $CP$. Here $SS_T$, $SS_E$, and $SS_M$ are sums of squares and can take only positive values, while $CP$ can be positive or negative . $CP$ represents the variation created (or lost) due to the correlation between the model variation. In the OLS case, $CP$ is zero for models with a constant, but this is not true for the general case. The following discussion illustrates why.

*Untransformed Empirical ANOVA for the OLS Case*

Assume $\widehat{\boldsymbol{Y}}_J$ is estimated from a linear model having the form:

$$\widehat{\boldsymbol{Y}}_J = \boldsymbol{X}\boldsymbol{b}_J \tag{5.14}$$

If $\boldsymbol{b}_J$ is the least squares solution of an OLS analysis, then by equation (5.14):

$$\begin{aligned} \widehat{\boldsymbol{Y}}_{OLS} = \boldsymbol{X}\boldsymbol{b}_{OLS} &= \boldsymbol{X}\boldsymbol{W}_{OLS}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\ &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{X}\boldsymbol{W}_{OLS}\boldsymbol{\varepsilon} \end{aligned} \tag{5.15}$$

Because $\boldsymbol{b}_{OLS}$ is the least squares estimator, it minimizes $SS_E$ from equation (5.13). Setting the first derivative of $SS_E$ equal to zero yields the OLS normal equation:

$$\boldsymbol{X}^T\left(\widetilde{\boldsymbol{Y}} - \boldsymbol{X}\boldsymbol{b}_{OLS}\right) = 0 \tag{5.16}$$

From equation (5.14), $\widehat{\boldsymbol{Y}}_J{}^T = \boldsymbol{b}_J{}^T\boldsymbol{X}^T$. Thus,

$$\boldsymbol{b}_{OLS}{}^T\boldsymbol{X}^T\left(\widetilde{\boldsymbol{Y}} - \boldsymbol{X}\boldsymbol{b}_{OLS}\right) = \widehat{\boldsymbol{Y}}_{OLS}{}^T\left(\widetilde{\boldsymbol{Y}} - \boldsymbol{X}\boldsymbol{b}_{OLS}\right) = 0 \tag{5.17}$$

Also,

$$\bar{y}\boldsymbol{X}^T\left(\widetilde{\boldsymbol{Y}} - \boldsymbol{X}\boldsymbol{b}_{OLS}\right) = 0 \tag{5.18}$$

Note that $\overline{\boldsymbol{Y}} = \bar{y}\boldsymbol{e}$. If $\boldsymbol{X}$ contains a column of ones, meaning that the model has a constant, equation (5.18) includes

$$\overline{\boldsymbol{Y}}^T\left(\widetilde{\boldsymbol{Y}} - \boldsymbol{X}\boldsymbol{b}_{OLS}\right) = 0 \tag{5.19}$$

Combining ( and (, and recalling the definition of $CP$ from (5.13), the $CP$ for the OLS case, $CP(OLS)$ is

$$\begin{aligned} CP(OLS) &= 2\left(\widetilde{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_{OLS}\right)^T\left(\widehat{\boldsymbol{Y}}_{OLS} - \overline{\boldsymbol{Y}}\right) \\ &= 2\left(\widetilde{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_{OLS}\right)^T\widehat{\boldsymbol{Y}}_{OLS} - 2\left(\widetilde{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_{OLS}\right)^T\overline{\boldsymbol{Y}} \\ &= 2\left(\widetilde{\boldsymbol{Y}} - \boldsymbol{X}\boldsymbol{b}_{OLS}\right)^T\widehat{\boldsymbol{Y}} - 2\left(\widetilde{\boldsymbol{Y}} - \boldsymbol{X}\boldsymbol{b}_{OLS}\right)^T\overline{\boldsymbol{Y}} \\ &= trace\left(2\left(\widetilde{\boldsymbol{Y}} - \boldsymbol{X}\boldsymbol{b}_{OLS}\right)^T\widehat{\boldsymbol{Y}}_{OLS}\right) - trace\left(2\left(\widetilde{\boldsymbol{Y}} - \boldsymbol{X}\boldsymbol{b}_{OLS}\right)^T\overline{\boldsymbol{Y}}\right) \\ &= 2trace\left(\widehat{\boldsymbol{Y}}_{OLS}{}^T\left(\widetilde{\boldsymbol{Y}} - \boldsymbol{X}\boldsymbol{b}_{OLS}\right)\right) - 2trace\left(\overline{\boldsymbol{Y}}^T\left(\widetilde{\boldsymbol{Y}} - \boldsymbol{X}\boldsymbol{b}_{OLS}\right)\right) \end{aligned} \tag{5.20}$$

$$CP(OLS) = 0$$

Thus, for the OLS case with a constant term, equation (5.13) becomes the classic relationship

$$SS_T(OLS) = SS_E(OLS) + SS_M(OLS) \tag{5.21}$$

The OLS empirical ANOVA table (Table 5.2) is often reported in regression studies [Draper and Smith, 1966]. The coefficient of determination, commonly

referred to as $R^2$, reports the fraction of variability which is explained by the model. $R^2$ is computed as one minus the ratio of $SS_E(OLS)$ and $SS_T(OLS)$.

$$R^2 = 1 - \frac{(\tilde{\boldsymbol{Y}}-\hat{\boldsymbol{Y}}_J)^T(\tilde{\boldsymbol{Y}}-\hat{\boldsymbol{Y}}_J)}{(\tilde{\boldsymbol{Y}}-\bar{\boldsymbol{Y}})^T(\tilde{\boldsymbol{Y}}-\bar{\boldsymbol{Y}})} = 1 - \frac{SS_E(OLS)}{SS_E(OLS)} \tag{5.22}$$

Table 5.2 is referred to as the OLS "empirical" ANOVA because it is computed directly from the data rather than estimated from the analysis. This distinction becomes important later in this section.

**Table 5.2:** OLS empirical ANOVA table

| Source | Sum of Squares | Degrees of Freedom |
|---|---|---|
| $SS_M(OLS)$ | $(\hat{\boldsymbol{Y}}-\bar{\boldsymbol{Y}})^T(\hat{\boldsymbol{Y}}-\bar{\boldsymbol{Y}})$ | k |
| $SS_E(OLS)$ | $(\tilde{\boldsymbol{Y}}-\hat{\boldsymbol{Y}})^T(\tilde{\boldsymbol{Y}}-\hat{\boldsymbol{Y}})$ | n-k-1 |
| $SS_T(OLS)$ | Sum of Above | n-1 |
| $R^2$ | $1 - SS_E(OLS)/SS_T(OLS)$ | |

*Untransformed empirical ANOVA for the General Case*

Now consider equation (5.13) for the more general case. In the WLS or GLS cases $CP$ does not become zero. This is because the analysis no longer attempts to minimize the sum squared errors, $SS_E$, from equation (5.13). Consider now the $J$ case, where $J$ = WLS or GLS. For the $J$ case, equation (5.15) becomes [Draper and Smith, 1966]:

$$\hat{\boldsymbol{Y}}_J = \boldsymbol{X}\boldsymbol{b}_J = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Lambda}_J{}^{-1}\boldsymbol{X})\boldsymbol{X}^T\boldsymbol{\Lambda}_J{}^{-1}\tilde{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{W}_J\tilde{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{W}_J(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}) \tag{5.23}$$

$$= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{X}\boldsymbol{W}_J\boldsymbol{\varepsilon}$$

Recall that for the general case, $E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T] = \boldsymbol{\Lambda}_J$. The sum of squared errors, $SS_E$, is no longer minimized by $\boldsymbol{b}_J$. Rather, the 'least squares' now refers to the sum of

squared weighted residuals, $\boldsymbol{\varepsilon}_*^T\boldsymbol{\varepsilon}_*$, which are minimized by $\boldsymbol{b}_J$ [Draper and Smith, 1966]:

$$\begin{aligned}\boldsymbol{\varepsilon}_*^T\boldsymbol{\varepsilon}_* &= \left(\widetilde{\boldsymbol{Y}} - \boldsymbol{X}\boldsymbol{b}_J\right)^T \boldsymbol{P}^T\boldsymbol{P}\left(\widetilde{\boldsymbol{Y}} - \boldsymbol{X}\boldsymbol{b}_J\right) \\ &= \left(\widetilde{\boldsymbol{Y}} - \boldsymbol{X}\boldsymbol{b}_J\right)^T \boldsymbol{\Lambda}_J^{-1}\left(\widetilde{\boldsymbol{Y}} - \boldsymbol{X}\boldsymbol{b}_J\right)\end{aligned} \tag{5.24}$$

where $\boldsymbol{\varepsilon}_*$ is a $k \times 1$ vector of transformed $\boldsymbol{\varepsilon}$, and $\boldsymbol{P}^T\boldsymbol{P} = \boldsymbol{\Lambda}_J^{-1}$. The $J$ case normal equation becomes:

$$\boldsymbol{X}^T\boldsymbol{\Lambda}_J^{-1}\left(\widetilde{\boldsymbol{Y}} - \boldsymbol{X}\boldsymbol{b}_J\right) = 0 \tag{5.25}$$

Thus for $J$ = WLS or GLS, $\widehat{\boldsymbol{Y}}_J^{\,T}\left(\widetilde{\boldsymbol{Y}} - \boldsymbol{X}\boldsymbol{b}_J\right) \neq 0$ and in general $\overline{\boldsymbol{Y}}^T\left(\widetilde{\boldsymbol{Y}} - \boldsymbol{X}\boldsymbol{b}_J\right) \neq 0$ and consequently $CP(J) \neq 0$. An exception is the trivial case that $\boldsymbol{\Lambda}_{\text{WLS}} = a\mathbf{I}$ for some constant $a$, which is actually an OLS analysis. Similarly, it can be shown that the $CP$ term for the GLS case, $CP(GLS)$, equals zero for the trivial case that all diagonal elements of $\boldsymbol{\Lambda}_{\text{GLS}}$ equal some constant $a$ and the off diagonals all equal some constant $b$.

**Table 5.3:** Empirical ANOVA table for WLS or GLS Regression

| Source | Sum of Squares |
|---|---|
| $SS_M(J)$ | $\left(\widehat{\boldsymbol{Y}}_J - \overline{\boldsymbol{Y}}\right)^T\left(\widehat{\boldsymbol{Y}}_J - \overline{\boldsymbol{Y}}\right)$ |
| $SS_E(J)$ | $\left(\widetilde{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_J\right)^T\left(\widetilde{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_J\right)$ |
| $CP(J)$ | $2\left(\widetilde{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_J\right)^T\left(\widehat{\boldsymbol{Y}}_J - \overline{\boldsymbol{Y}}\right)$ |
| $SS_T$ | Sum of Above |
| $R^2$ | $1 - SS_E(J)/SS_T(J)$ |

Interestingly, $CP(WLS) \neq 0$ shows that the correlation between the residuals and the model variation occurs even when the observations are not correlated. Thus, the OLS ANOVA division of variability does not work for the GLS case *or* the WLS

case, because $SS_T \neq SS_M + SS_E$. The WLS or GLS empirical ANOVA table should include the $CP$ term if all of the sample variability is to be accounted for, as in Table 5.3.

Several problems arise in Table 5.3. Because $CP$ can take negative values, $R^2$ is no longer bounded by [0,1]. This is problematic for our understanding of $R^2$: $R^2 > 1$ means that the model is explaining more variability than exists in the data. For this reason the traditional $R^2$ is not a reliable means to compare different models for the GLS *or* WLS cases [Buse 1973; Blomquist, 1980]. The presence of the *CP(J)* term, that can be positive or negative, results in an ANOVA table that is much harder to interpret. Another issue with Table 5.3 is that it is no longer clear how many degrees of freedom each quantity reflects, and a degrees of freedom column is not included with the table.

***Section 5.2.1.2 Transformed Empirical ANOVA***

This section explores several empirical approaches to ANOVA for GLS which divide variability between sources in a transformed space rather than in the original problem space. As described, these methods originate from the interpretation of WLS and GLS as OLS in a transformed space where the errors are homoscedastic and uncorrelated. This can raise new concerns.

*Draper and Smith [1966] ANOVA*

To structure an ANOVA for the WLS or GLS cases, Draper and Smith [1966] propose transforming the data and residuals to a space where $CP = 0$. Consider the $J$ case for *J=WLS* or *GLS*. Recall that $\boldsymbol{P}^T\boldsymbol{P} = \boldsymbol{\Lambda}_J^{-1}$. Pre-multiply both sides of equation (5.6) by $\boldsymbol{P}^T$:

$$\boldsymbol{P}^T\tilde{\boldsymbol{Y}} = \boldsymbol{P}^T\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{P}^T\boldsymbol{\varepsilon} \tag{5.26}$$

$$\tilde{\boldsymbol{Z}} = \boldsymbol{Q}\boldsymbol{\beta} + \boldsymbol{f}$$

Consider the OLS model in the transformed space:

$$\hat{\boldsymbol{Z}} = \boldsymbol{Q}\boldsymbol{b}_{Z,OLS} \tag{5.27}$$

The $J$ case parameter estimators, $\boldsymbol{b}_J$, are the OLS parameter estimators for the transformed space, $\boldsymbol{b}_{Z,OLS}$ [Draper and Smith, 1966; Greene, 2008]:

$$\boldsymbol{b}_{Z,OLS} = (\boldsymbol{Q}^T\boldsymbol{Q})^{-1}\boldsymbol{Q}^T\tilde{\boldsymbol{Z}} = \left(\boldsymbol{X}^T\boldsymbol{\Lambda}_J^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\tilde{\boldsymbol{Y}} = \boldsymbol{b}_J \tag{5.28}$$

The transformed total sum of squares, $\widetilde{SS}_T$ is provided by:

$$\begin{aligned}\widetilde{SS}_T(J) &= \left(\tilde{\boldsymbol{Z}} - \overline{\boldsymbol{Z}}\right)^T\left(\tilde{\boldsymbol{Z}} - \overline{\boldsymbol{Z}}\right) \\ &= \left(\tilde{\boldsymbol{Z}} - \hat{\boldsymbol{Z}}\right)^T\left(\tilde{\boldsymbol{Z}} - \hat{\boldsymbol{Z}}\right) + \left(\hat{\boldsymbol{Z}} - \overline{\boldsymbol{Z}}\right)^T\left(\hat{\boldsymbol{Z}} - \overline{\boldsymbol{Z}}\right) - 2\left(\tilde{\boldsymbol{Z}} - \hat{\boldsymbol{Z}}\right)^T\left(\hat{\boldsymbol{Z}} - \overline{\boldsymbol{Z}}\right) \\ \widetilde{SS}_T(J) &= \widetilde{SS}_E(J) + \widetilde{SS}_M(J) - 2\widetilde{CP}(J)\end{aligned} \tag{5.29}$$

where $\overline{\boldsymbol{Z}}$ is a $n \times 1$ vector containing the sample mean of the $\tilde{Z}$. Here, $\widetilde{SS}_T$ is divided into three components: sum of squares due to error, $\widetilde{SS}_E$, and sum of squares due to the model, $\widetilde{SS}_M$, and a cross-product term $\widetilde{CP}$. Draper and Smith [1966, pg. 78-79] claim that $\widetilde{CP}(J) = 0$ and thus a natural extension is a transformed $R^2$, $\tilde{R}^2$, defined as one minus the ratio of $\widetilde{SS}_E$ and $\widetilde{SS}_T$. However, Buse [1973] correctly points out that $\widetilde{CP}(J) \neq 0$ in every case, as there is no guarantee that the transformed independent variables $\boldsymbol{Q}$ contain a constant (column of ones) and in general it will not. Table 5.4 reports the transformed WLS or GLS empirical ANOVA. Again note that this is an empirical table because it is computed directly from the data.

It is not clear that this table provides any additional value over the empirical ANOVA table (Table 5.3): its prime motivation was that $\widetilde{CP}(J)$ disappears, which turns

out not to be true. Furthermore, the transformed empirical ANOVA reports variation in a space which is solely a computational construct and maybe of no interest or relevance to the analyst [Greene, 2008, 156].

**Table 5.4:** Empirical ANOVA for a Transformed WLS or GLS Analysis

| Source | Sum of Squares |
|---|---|
| $\widetilde{SS}_M(J)$ | $\left(\hat{\boldsymbol{Z}}-\overline{\boldsymbol{Z}}\right)^T\left(\hat{\boldsymbol{Z}}-\overline{\boldsymbol{Z}}\right)$ |
| $\widetilde{SS}_E(J)$ | $\left(\tilde{\boldsymbol{Z}}-\hat{\boldsymbol{Z}}\right)^T\left(\tilde{\boldsymbol{Z}}-\hat{\boldsymbol{Z}}\right)$ |
| $\widetilde{CP}(J)$ | $-2\left(\tilde{\boldsymbol{Z}}-\hat{\boldsymbol{Z}}\right)^T\left(\hat{\boldsymbol{Z}}-\overline{\boldsymbol{Z}}\right)$ |
| $\widetilde{SS}_T(J)$ | Sum of Above |
| $\tilde{R}^2$ | $1-\widetilde{SS}_E(J)/\widetilde{SS}_T(J)$ |

*Buse [1973] ANOVA*

Buse [1973] recognizes the problem with Table 5.4, and devised an alternative empirical $R^2$ for the GLS case which is bounded on [0,1]. The following discussion describes the ANOVA which is implicitly suggested by his empirical $R^2$. Buse [1973] shows that if the model includes a constant,

$$\left(\tilde{\boldsymbol{Y}}-\bar{y}_J\boldsymbol{e}\right)^T\boldsymbol{\Lambda}_J^{-1}\left(\tilde{\boldsymbol{Y}}-\bar{y}_J\boldsymbol{e}\right) \tag{5.30}$$

$$=\left(\hat{\boldsymbol{Y}}_J-\bar{y}_J\boldsymbol{e}\right)^T\boldsymbol{\Lambda}_J^{-1}\left(\hat{\boldsymbol{Y}}_J-\bar{y}_J\boldsymbol{e}\right)+\left(\tilde{\boldsymbol{Y}}-\hat{\boldsymbol{Y}}_J\right)^T\boldsymbol{\Lambda}_J^{-1}\left(\tilde{\boldsymbol{Y}}-\hat{\boldsymbol{Y}}_J\right)$$

where $\bar{y}_J$ is the constant model for the $J$ case defined:

$$\bar{y}_J=\left(\boldsymbol{e}^T\boldsymbol{\Lambda}_J^{-1}\boldsymbol{e}\right)^{-1}\boldsymbol{e}^T\boldsymbol{\Lambda}_J^{-1}\tilde{\boldsymbol{Y}} \tag{5.31}$$

The term on the left-hand-side of equation (5.30) describes the total generalized sum of squares about the weighted mean. The first term on the right-hand-side of equation (5.30) describes the generalized variability about the weighted mean explained by the model, and the second term describes the generalized

variability about the model. A generalized empirical ANOVA can be based on equation (5.30), as shown in Table 5.5.

**Table 5.5:** Generalized WLS or GLS empirical ANOVA

| Source | Sum of Squares |
|---|---|
| $SS_{M,Buse}(J)$ | $\left(\hat{\boldsymbol{Y}}_J - \bar{y}_J \boldsymbol{e}\right)^T \boldsymbol{\Lambda}_J^{-1} \left(\hat{\boldsymbol{Y}}_J - \bar{y}_J \boldsymbol{e}\right)$ |
| $SS_{E,Buse}(J)$ | $\left(\tilde{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_J\right)^T \boldsymbol{\Lambda}_J^{-1} \left(\tilde{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_J\right)$ |
| $SS_{T,Buse}(J)$ | Sum of Above |
| ${R_{Buse}}^2$ | $1 - SS_{E,Buse}(J)/SS_{T,Buse}(J)$ |

Unlike $\tilde{R}^2$, ${R_{Buse}}^2$ is bounded on [0,1]. However, like the transformed empirical ANOVA in Table 5.4, this analysis is still concerned with deviations in transformed space, which is purely a computational construct and might be of no interest to the analyst. In particular, for the simpler WLS case where the variances on the diagonal of $\Lambda_J$ vary widely, the weights places in Table 5.4 on the deviations of model predictions from the average, and on different errors, will also vary widely. So what do such sum of squares mean?

#### *Section 5.2.1.3 A note on other empirical $R^2$ Statistics*

Several empirical ANOVA and associated coefficients of determination have been described in this section. Others are described by La Du and Tanaka [1989]. There seems to be no universally accepted pseudo ANOVA procedure for GLS analyses. Greene [2008] prefers the definition of $R^2$ reported in Table 5.3, and thus implicitly prefers the empirical ANOVA over the transformed empirical ANOVA.

Blomquist [1980] recommends the Baxter and Cragg [1970] "pseudo $R^2$" based on likelihood ratios. The Baxter and Cragg [1970] "pseudo $R^2$" is computed as

$$R_P^2 = \frac{1 - \exp\left(\frac{2(L_\omega - L_\Omega)}{n}\right)}{1 - \exp\left(\frac{2(L_\omega - L_\Omega^{max})}{n}\right)} \tag{5.32}$$

where $L_\omega$ is the log maximum likelihood function value when only a constant is used, and $L_\Omega$ is the log maximum likelihood function value for the final model and $L_\Omega^{max}$ is the log maximum achievable value of the likelihood function. Similar "pseudo $R^2$" are recommended by Menard [2000] for logistic regression and are commonly applied in maximum likelihood discrete regression, wherein the dependent variable is discrete [Long, 1997]. These "pseudo $R^2$" are typically based on some deviance statistic, which is a measure of the distance between $L_\Omega$ and $L_\Omega^{max}$. A natural "pseudo $R^2$", based on this deviance statistic[Cameron and Trivedi, 1998]

$${R_{DEV}}^2 = 1 - \frac{(L_\Omega - L_\Omega^{max})}{(L_\omega - L_\Omega^{max})} \tag{5.33}$$

Liu et al. [2005] propose a differed "pseudo $R^2$" for negative binomial regression:

$$R_\alpha^2 = 1 - \frac{\alpha}{\alpha_0} \tag{5.34}$$

where $\alpha$ and $\alpha_0$ are the variances of the gamma-distributed error for fitted final model and the constant model respectively. This statistic can be seen as directly related to the "pseudo $R^2$" statistic proposed for Bayesian GLS by Gruber et al. [2007]. In a different vein, Gelman and Pardoe [2006] reviews Bayesian $R^2$ for hierarchical models (related to ANOVA "type I").

Buse [1973] describes the various shortcomings of the $R^2$ as a standalone statistic, and recommends the analyst use caution. Similarly, Jarrett [1974] finds the

$R^2$ is useful, but should be paired with other analyses, perhaps graphical, to fully understand a model performance.

***Section 5.2.2: Theoretical ANOVA approaches***

The previous section described several empirical approaches to ANOVA for WLS and GLS analyses. An additional concern in the Stedinger and Tasker [1985] hydrologic regression framework is the division of error between model error and sampling error. This division is not easily estimated by empirical methods, so theoretical methods based on the variance and correlation structure of the data have been used. This section describes the development of the pseudo ANOVA for the Stedinger-Tasker model and compares it to empirical approaches.

In the Stedinger and Tasker [1985] hydrologic regression framework, the error is composed of model error and sampling error. In this framework, equation (5.6) is expanded to

$$\widetilde{\boldsymbol{Y}} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} = \boldsymbol{X\beta} + \boldsymbol{\delta} + \boldsymbol{\eta} \tag{5.35}$$

where $\boldsymbol{\delta}$ is a $n \times 1$ vector of model errors and $\boldsymbol{\eta}$ is a $n \times 1$ vector of sampling errors. Model error is due to the use of an imperfect model. Sampling error is due to the use of finite records to estimate each of the $y$-observations. It is assumed that both error types have zero mean, i.e. $E[\boldsymbol{\delta}] = E[\boldsymbol{\eta}] = \mathbf{0}$. The covariance matrix of $\boldsymbol{\varepsilon}$, $E[\boldsymbol{\varepsilon\varepsilon}^T]$, is given by:

$$E[\boldsymbol{\varepsilon\varepsilon}^T] = E[\boldsymbol{\delta\delta}^T] + E[\boldsymbol{\eta\eta}^T] = \mathbf{I}{\sigma_\delta}^2(k) + \boldsymbol{\Sigma}\left(\widetilde{\boldsymbol{Y}}\right) = \boldsymbol{\Lambda} \tag{5.36}$$

where $\boldsymbol{\Sigma}\left(\widetilde{\boldsymbol{Y}}\right)$ is the sampling error covariance matrix, and ${\sigma_\delta}^2(k)$ is the model error variance of the model with $k-1$ explanatory variables. Let ${\sigma_\delta}^2(0)$ be the model error variance of the constant model. Note that the model errors are

uncorrelated and are assumed to have constant variance for every observation. In contrast, the sampling errors can be correlated between sites and do not have equal variance.

Even a perfect model [$\sigma_\delta{}^2(k) = 0$] cannot explain sampling error, so using $SS_E$ in the computation of $R^2$ in Table 5.3 would not provide the desired insight. Instead, Reis [2005] proposes using the Bayesian estimate of the model error variance of the fitted model, to estimate the unexplained variation in the true $\boldsymbol{Y}$. This follows by noting that

$$E[\boldsymbol{\delta}^T\boldsymbol{\delta}] = trace\left[E[\boldsymbol{\delta}\boldsymbol{\delta}^T]\right] = nE[\sigma_\delta^2(k)] = n\hat{\sigma}_\delta^2(k) \tag{5.37}$$

Thus, the expected model error sum-of-squares is $n\hat{\sigma}_\delta^2(k)$, where $\hat{\sigma}_\delta^2(k)$ is the Bayesian estimate of $E[\sigma_\delta^2(k)]$.

To estimate the variability in the true skew explained by the model, Reis [2005] recommends the empirical estimator:

$$\sum\left(\hat{y}_i - \bar{y}_J\right)^2 \tag{5.38}$$

where $\bar{y}_J$is the constant model from either a $J$=WLS or GLS analysis. The estimate of the total variation in the true skew is then:

$$n\hat{\sigma}_\delta{}^2(k) + \sum\left(\hat{y}_i - \bar{y}_J\right)^2 \tag{5.39}$$

Noting that $R^2$ is one minus the ratio of the model error and the total variability, Reis [2005] proposes the pseudo $R^2$:

$$R_J{}^2(pseudo) = 1 - \frac{n\hat{\sigma}_\delta{}^2(k)}{n\hat{\sigma}_\delta{}^2(k) + \sum\left(\hat{y}_i - \bar{y}_J\right)^2} = \frac{\sum\left(\hat{y}_i - \bar{y}_J\right)^2}{n\hat{\sigma}_\delta{}^2(k) + \sum\left(\hat{y}_i - \bar{y}_J\right)^2} \tag{5.40}$$

This implicitly suggests the simple pseudo ANOVA in Table 5.6; however Reis [2005] did not recommend such an ANOVA. Sample error could also be added to this table.

**Table 5.6:** Pseudo ANOVA table based on Reis [2005] pseudo $R^2$, ${R_J}^2(pseudo)$

| Source | Sum of Squares |
|---|---|
| Model | $\sum(\hat{y}_i - \bar{y}_J)^2$ |
| Model Error | $n\hat{\sigma}_\delta^2(k)$ |
| Variability in $y$ | Sum of Above |
| ${R_J}^2(pseudo)$ | $1 - \frac{n\hat{\sigma}_\delta^2(k)}{n\hat{\sigma}_\delta^2(k) + \sum(\hat{y}_i - \bar{y}_J)^2}$ |

Because ${R_J}^2(pseudo)$ considers only true $y$ variation, this implicit table considers only true $y$ variation. This is an advantage because it recognizes that even a perfect model will not explain all of the variability in the observed data. Because the estimate of the total variability is based on the estimate of $\hat{\sigma}_\delta^2(k)$ and an empirical estimate of model variability, it will change depending on the postulated model. This is problematic: the variability in the data should not depend on the selected model. This highlights a flaw in the Reis [2005] ${R_J}^2(pseudo)$: two models with the same model error variance could have different ${R_J}^2(pseudo)$ values if the models have different $\sum(\hat{y}_i - \bar{y}_J)^2$. Thus, this ${R_J}^2(pseudo)$ may have trouble comparing competing models, and by extension Table 5.6 is a flawed ANOVA for comparing the performance of different models. On the other hand, even with OLS, models with larger $R^2$ values need not have statistically significant parameters in comparison to models with smaller $R^2$ values, and $R^2$ is known to increase whenever another

explanatory variable is added. Clearly $R^2$ values should not be the primary criteria for model selection.

Gruber et al. [2007] propose a theoretical pseudo ANOVA and pseudo $R^2$ which is based entirely on estimates of the sampling error and model error variances, rather than computed from the data. Like Reis [2005], Gruber et al. note that the expected value of the sum of squares due to model error would be

$$SS(Model\ Error) = E[\boldsymbol{\delta}^T\boldsymbol{\delta}] = trace\left[E[\boldsymbol{\delta}\boldsymbol{\delta}^T]\right] = n\hat{\sigma}_\delta^2(k) \tag{5.41}$$

Similarly, the expected value of the sum of squares due to sampling error in the observations is

$$SS(Sample\ Error) = E[\boldsymbol{\eta}^T\boldsymbol{\eta}] = trace\left[E[\boldsymbol{\eta}\boldsymbol{\eta}^T]\right] = trace\left[\boldsymbol{\Sigma}(\tilde{\boldsymbol{Y}})\right] = \sum_{i=1}^{n} var(\tilde{y}_i) \tag{5.42}$$

The constant model can explain no variability in the data. Thus the model error variance of the constant model, $\sigma_\delta^2(0)$, describes the variability of the true $\boldsymbol{Y}$ by correctly deducting sampling error variability. Thus, an estimate of the total sum of squares in the true skew is given by:

$$SS(True\ Skew) = E\left[\boldsymbol{\delta}_0{}^T\boldsymbol{\delta}_0\right] = trace\left[E\left[\boldsymbol{\delta}_0\boldsymbol{\delta}_0{}^T\right]\right] = n\hat{\sigma}_\delta^2(0) \tag{5.43}$$

where $\boldsymbol{\delta}_0$ is a $n \times 1$ vector of model errors for the constant model. Gruber et al. [2007] then propose a pseudo $R^2$, ${R_\delta}^2$, computed as:

$${R_\delta}^2 = 1 - \frac{n\hat{\sigma}_\delta^2(k)}{n\hat{\sigma}_\delta^2(0)} = 1 - \frac{\hat{\sigma}_\delta^2(k)}{\hat{\sigma}_\delta^2(0)} \tag{5.44}$$

Like the Reis [2005] ${R_J}^2(pseudo)$, the value of ${R_\delta}^2$ is bounded on the interval [0,1]. A model with no model error, i.e. $\hat{\sigma}_\delta^2(k) = 0$, will have ${R_\delta}^2 = 1$. A model which explains no variability, i.e. $\hat{\sigma}_\delta^2(k) = \hat{\sigma}_\delta^2(0)$, will have ${R_\delta}^2 = 0$. Also, like the

Reis [2005] pseudo $R^2$, ${R_\delta}^2$ only incorporates model error variability. This is crucial because even a perfect model will not explain sampling error variability. A natural extension of ${R_\delta}^2$ is their pseudo ANOVA table (Table 5.7), which reports the breakdown of variability between model error (equation(5.41)), sampling error (equation (5.42)), and variability explained by the model. This final term is computed by subtracting the model error sum of squares ($SS(Model\ Error)$) from the total sum of squares in the true $\boldsymbol{Y}$ ($SS(True\ Skew)$):

$$SS(Model) = SS(True\ Skew) - SS(Model\ Error) = n\hat{\sigma}_\delta^2(0) - n\hat{\sigma}_\delta^2(k) \qquad (5.45)$$

$$= n\left[\hat{\sigma}_\delta^2(0) - \hat{\sigma}_\delta^2(k)\right]$$

The pseudo ANOVA proposed by Gruber et al. [2007] (Table 5.7) is an improvement over Table 5.6 because it includes the sampling error sum of squares. One way to determine if a WLS or GLS analysis is preferred over an OLS analysis is to compare the relative magnitude of the model error variability and sampling error variability. Table 5.7 makes this possible. Another feature is that the estimate of the total sum of squares, $SS(Total)$, does not depend on the selected model. A related feature is that ${R_\delta}^2$ does not depend the empirical variation of the model about a mean, and consequently two models with the same model error variance will have the same ${R_\delta}^2$. This is an improvement over the Reis [2005] ${R_J}^2(pseudo)$.

A key advantage of the Table 5.7 over the empirical ANOVA in Table 5.3 is that Table 5.7 divides the error variability into sampling error and model error. A second advantage is that sampling errors are removed in the computation of ${R_\delta}^2$, reflecting the fact that even a perfect model cannot explain all of the variability in a

data set. For conciseness in the following discussion, Table 5.7 will be referred to as the B-GLS ANOVA, for Bayesian GLS ANOVA.

**Table 5.7:** Pseudo ANOVA table for WLS or GLS regression [Gruber et al., 2007]

| Source | Sum of Squares | Degrees-of-freedom |
|---|---|---|
| $SS(Model)$ | $n\left[\hat{\sigma}_\delta^2(0) - \hat{\sigma}_\delta^2(k)\right]$ | $k$ |
| $SS(Model\ Error)$ | $n\hat{\sigma}_\delta^2(k)$ | $n-k-1$ |
| $SS(Sample\ Error)$ | $\sum_{i=1}^{n} var(\tilde{y}_i)$ | $n$ |
| $SS(Total)$ | Sum of Above | $2n-1$ |
| ${R_\delta}^2$ | $1 - \frac{\hat{\sigma}_\delta^2(k)}{\hat{\sigma}_\delta^2(0)}$ | |

A criticism of Table 5.7 is that it incorrectly assigns degrees of freedom to each quantity. If the sampling errors are positively correlated, then the effective number of independent observations is potential much smaller than $n$ (see Stedinger [1983] for a discussion of this issue). Furthermore, the estimate of the variation in the true $\boldsymbol{Y}$ is correlated with the sampling errors (i.e. recall the interpretation of the $CP(J)$ term from equation (5.13)). Finally, if $\hat{\sigma}_\delta^2(k)$ and $\hat{\sigma}_\delta^2(0)$ are estimated with a Bayesian analysis, then the prior distribution of $\delta$ affects the effective degrees of freedom. This last point is most important for small sample sizes, with large correlation. Thus it may be best to delete the degrees of freedom column.

It is important to observe that the pseudo ANOVA is based solely on the theoretical breakdown of error among different sources and does not depend solely on the observations. A consequence of this is that $SS(Total) \neq E[SS_T]$ from equation (5.47), and thus it is difficult to compare empirical ANOVA results with theoretical

pseudo ANOVA results. This in itself is not necessarily troubling, but should be understood. The following section explores an alternative theoretical ANOVA based on $E[SS_T]$.

***Section 5.2.3: New Pseudo ANOVA***

The previous section described theoretical pseudo ANOVA approaches proposed by Reis [2005] and Gruber et al. [2007]. This section explores an alternative ANOVA for the Stedinger and Tasker [1985] regression framework based on the theoretical value of $E[SS_T]$. Recall from equation (5.13) that

$$SS_T = SS_E(J) + SS_M(J) + 2CP(J)$$

where

$$\begin{aligned} SS_T &= \left(\tilde{\boldsymbol{Y}} - \overline{\boldsymbol{Y}}\right)^T \left(\tilde{\boldsymbol{Y}} - \overline{\boldsymbol{Y}}\right) \\ SS_E(J) &= \left(\tilde{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_J\right)^T \left(\tilde{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_J\right) \\ SS_M(J) &= \left(\hat{\boldsymbol{Y}}_J - \overline{\boldsymbol{Y}}\right)^T \left(\hat{\boldsymbol{Y}}_J - \overline{\boldsymbol{Y}}\right) \\ CP(J) &= \left(\tilde{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_J\right)^T \left(\hat{\boldsymbol{Y}}_J - \overline{\boldsymbol{Y}}\right) \end{aligned} \tag{5.46}$$

and $J$ indicates whether an OLS, WLS, or GLS analysis is used to estimate the regression model parameters, and $SS_T$, $SS_E$, $SS_M$, and $CP$ are the total observed sum-of-squares, observed error sum of squares, observed model sum of squares, and observed cross-product term respectively. A new pseudo ANOVA can be based on a estimate of $E[SS_T]$. First, consider the case that a $J = WLS$ or $GLS$ analysis is used to estimate both the model parameters and their precision. Let $\boldsymbol{N}$ be a $n \times n$ matrix with each element equal $1/n$ . The derivation in the appendix shows that

$$E[SS_T] = (\boldsymbol{X\beta})^T(\boldsymbol{I}-\boldsymbol{N})\boldsymbol{X\beta} + SS(Model\ Error) + SS(Sampling\ Error) - trace(\boldsymbol{N\Lambda}_J)$$

$$= (\boldsymbol{X\beta})^T(\boldsymbol{I}-\boldsymbol{N})\boldsymbol{X\beta} + (n-1)\hat{\sigma}_\delta^2(k) + (n-1)\sum_{i=1}^{n}\frac{var(\tilde{y}_i)}{n} - \frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}\boldsymbol{\Lambda}_J(i,j) \tag{5.47}$$

$$E[SS_T] = SS_*(Model) + SS_*(Model\ Error) + SS_*(Sampling\ Error) - \rho\text{- Correction}$$

The first line in equation (5.47) relates $E[SS_T]$ to the Gruber et al. [2007] pseudo ANOVA (Table 5.7). The second and third terms explain the variation due to model and sampling errors respectively, and are identical to those proposed by Gruber et al. [2007] (see Table 5.7). The first term describes the variation of the true model about its mean, and the fourth term is a correction for the errors in computing the sample mean.

An new pseudo ANOVA table can be based on the final line in equation (5.47). The first term on the last line of equation (5.47) reports the variation of the true model about its mean. The second and third terms describe the variation due to model error, and sampling error respectively. The last term is a correction for the correlation of the sampling errors. Unfortunately, $\boldsymbol{\beta}$ is unknown. If $\boldsymbol{b}_J$ is substituted for $\boldsymbol{\beta}$, the first term in equation (5.47) is similar to the term proposed by Reis [2005] to estimate the variation explained by the model (see equation (5.38)). Table 5.9 is an alternative pseudo ANOVA based on equation (5.47).

**Table 5.8:** New Pseudo ANOVA for WLS or GLS regression based on the Stedinger and Tasker [1985] regression framework, and the estimate of $\boldsymbol{E[SS_T]}$

| Source | Sum of Squares |
|---|---|
| $SS_*(Model)$ | $\widehat{\boldsymbol{Y}}_J^T\widehat{\boldsymbol{Y}}_J - \widehat{\boldsymbol{Y}}_J^T\boldsymbol{N}\widehat{\boldsymbol{Y}}_J$ |
| $SS_*(Model\ Error)$ | $(n-1)\hat{\sigma}_\delta^2(k)$ |
| $SS_*(Sample\ Error)$ | $(n-1)\sum_{i=1}^{n}\frac{var(\tilde{y}_i)}{n}$ |
| $\rho$-Correction | $-\frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}\boldsymbol{\Lambda}_J(i,j)$ |
| $SS_*(Total)$ | Sum of Above |

The pseudo ANOVA in Table 5.8 is appealing because it is a theoretical ANOVA based on $E[SS_T]$. The appendix of this chapter contains a derivation of an expression for $E[SS_T]$ for a hybrid WLS/GLS analysis, such as the one used in Chapter 4. In a WLS/GLS analysis, WLS is used to estimate the model parameters, and GLS is used to estimate the precision of the model. Table 5.9 contains an adapted alternative B-GLS pseudo ANOVA for the WLS/GLS case based on $E[SS_T]$.

**Table 5.9:** New Pseudo ANOVA for Hybrid WLS/GLS regression based on the Stedinger and Tasker [1985] regression framework, and the estimate of $\boldsymbol{E[SS_T]}$

| Source | Sum of Squares |
|---|---|
| $SS_*(Model)$ | $\widehat{\boldsymbol{Y}}_{WLS}^T\widehat{\boldsymbol{Y}}_{WLS} - \widehat{\boldsymbol{Y}}_{WLS}^T\boldsymbol{N}\widehat{\boldsymbol{Y}}_{WLS}$ |
| $SS_*(Model\ Error)$ | $(n-1)\hat{\sigma}_{\delta,GLS}^2(k)$ |
| $SS_*(Sample\ Error)$ | $(n-1)\sum_{i=1}^{n}\frac{var(\tilde{y}_i)}{n}$ |
| $\rho$-Correction | $-\frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}\boldsymbol{\Lambda}_{GLS}(i,j)$ |
| $SS_*(Total)$ | Sum of Above |

$SS_T$ computes variability about the sample mean. Because the sampling errors are heteroscedastic, it is unclear why the sample mean is appropriate. Buse [1973] addresses this by computing the variation about $\bar{y}_J$ rather than $\bar{y}$. Reis [2005] also takes this approach, estimating the variability explained by the model by the sum of squared variations about the constant model. Adopting this idea, define a new total sum of squares, computed about the constant WLS or GLS model:

$$SS_{T,J} = \left(\tilde{\boldsymbol{Y}} - \bar{\boldsymbol{Y}}_J\right)^T \left(\tilde{\boldsymbol{Y}} - \bar{\boldsymbol{Y}}_J\right) \tag{5.48}$$

where $\bar{\boldsymbol{Y}}_J$ is an $n \times 1$ vector containing the constant model for the $J$ case for $J = GLS$ or $WLS$. Recall the definition of $\boldsymbol{W}_J$ in equation (5.11). Let $\boldsymbol{C}_J$ be the corresponding weight vector for the constant model, for the $J$ case:

$$\boldsymbol{C}_J = \left(\boldsymbol{e}^T \boldsymbol{\Lambda}_J^{-1} \boldsymbol{e}\right)^{-1} \boldsymbol{e}^T \boldsymbol{\Lambda}_J^{-1} \tag{5.49}$$

where $\boldsymbol{e}$ is a $n \times 1$ vector of ones. Let $\boldsymbol{F}$ be a $n \times n$ matrix containing $\boldsymbol{C}_J$ on each row. Substituting $\boldsymbol{F}$ for $\boldsymbol{N}$ in the derivation of $E[SS_T]$, it can be shown that

$$\begin{aligned} E[SS_{T,J}] = {} & (\boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{I} - \boldsymbol{F}) \boldsymbol{X}\boldsymbol{\beta} + (n-1)\hat{\sigma}_\delta^2(k) + \sum_{i=1}^{n} \left(1 - c_J(i)\right) var(\tilde{y}_i) \\ & - \sum_{i=1}^{n} c_J(i) \sum_{j \neq i} \rho_{ij} \left(var(\tilde{y}_i)\right)^{0.5} \left(var(\tilde{y}_j)\right)^{0.5} \end{aligned} \tag{5.50}$$

where $c_J(i)$ is the $i^{th}$ element of $\boldsymbol{C}_J$. If $\boldsymbol{b}_J$ is substituted for $\boldsymbol{\beta}$, the first term in equation (5.50) is the nearly the same as the term proposed by Reis [2005] to describe the variation explained by the model. Table 5.9 contains a pseudo ANOVA based on equation (5.50).

Table 5.8, Table 5.9, and Table 5.10 have theoretical appeal. Table 5.8 is a theoretical extension of the OLS empirical ANOVA and matches $E[SS_T]$ in expectation. In that vein, Table 5.10 is divorced somewhat from the OLS empirical ANOVA because it is based on a computation of total variation about a weighted average. On the other hand, Table 5.10 correctly recognizes the heteroscedascity of the observations, and is more consistent with past GLS ANOVA efforts by Buse[1973] and Reis [2005].

**Table 5.10:** Alternative Pseudo ANOVA for WLS or GLS regression based on the Stedinger and Tasker [1985] regression framework, and the estimate of $\boldsymbol{E[SS_{T,J}]}$

| Source | Sum of Squares |
|---|---|
| $SS_{T,J}(Model)$ | $\widehat{\boldsymbol{Y}}_J^{\,T}\widehat{\boldsymbol{Y}}_J - \widehat{\boldsymbol{Y}}_J^{T}\boldsymbol{F}\widehat{\boldsymbol{Y}}_J$ |
| $SS_J(Model\ Error)$ | $(n-1)\hat{\sigma}_\delta^2(k)$ |
| $SS_J(Sample\ Error)$ | $\sum_{i=1}^{n}\left(1-c_J(i)\right)var(\tilde{y}_i)$ |
| $\rho$-Correction | $-\sum_{i=1}^{n}c_J(i)\sum_{j\neq i}\rho_{ij}\left(var(\tilde{y}_i)\right)^{0.5}\left(var(\tilde{y}_j)\right)^{0.5}$ |
| $E[SS_{T,J}]$ | Sum of Above |

There are several advantages of three new pseudo ANOVAs over that proposed by Gruber et al. [2007]. First, Table 5.8 is based directly on an estimator of $E[SS_T]$, and as such is much more closely related to the traditional empirical OLS ANOVA (Table 5.2). In fact the pseudo ANOVA in Table 5.8 can be seen as an extension of the traditional empirical ANOVA, wherein we attempt to estimate $E[SS_T]$ for a random sample rather than directly for the data. In contrast the B-GLS ANOVA does not attempt to match the traditional empirical ANOVA, but rather to

explain the implications of the fitted model and the contributions of different sources of variation to a random observation.

The new tables include all the terms which compose $E[SS_T]$ whereas the pseudo ANOVA in Table 5.7 admittedly neglects a correction for the correlation of the sampling errors. The alternative tables are perhaps most instructive in indicating how this term might be constructed.

Like the hybrid theoretical/empirical ANOVA implicitly suggested by Reis [2005]'s $R^2$, $R_f^2(pseudo)$, a troubling aspect of the two alternative pseudo ANOVA is that the estimate of the total variability in the data will depend on the choice of fitted model. This is clearly problematic, but the extent of the problem is not clear. An increase in $SS_J(Model)$ or $SS_*(Model)$ will be accompanied by a corresponding decrease in $\hat{\sigma}_\delta^2(k)$, but it is unlikely that these will perfectly compensate. The following section explores the extent of this problem through numerical examples.

***Section 5.2.4: Examples and Discussion***

The previous sections describe several conceptual frameworks for a GLS ANOVA, including procedures based upon both empirical and theoretical computations. In particular, Section 5.2.3 develops a new pseudo ANOVA table that approximates the expectation of the total sum of squares $SS_T$. This section explores the relative merits of the alternative pseudo ANOVA approaches using numerical examples to illustrate the differences. In particular, three ANOVA computations are compared: an empirical GLS ANOVA, the B-GLS pseudo ANOVA, and a new alternative B-GLS pseudo ANOVA based on $E[SS_T]$. A particular concern is the relative advantages of the two pseudo ANOVAs. The empirical GLS ANOVA is

included because it is closely related to the new pseudo ANOVA; unfortunately empirical GLS ANOVA procedures have no good way to provide out the average sampling variance, and thus are unable to address the relative importance of that critical source of variability.

The breakdown of the sum of squares for each ANOVA computation is reported for three fitted models: a constant model, a linear elevation model, and the final non-linear elevation model reported in Chapter 4. As discussed earlier, a feature of the new pseudo ANOVA is that the estimate of the total sum of squares depends on the fitted model, but it is not clear to what extent this is true.

The three ANOVA computations for each of the three fitted models are applied for two data sets: the 1-Day and 30-Day California skew data from Chapter 4 (see Table 5.11, Table 5.12, and Table 5.13). Sampling errors are fairly consistent between the two data sets corresponding to different durations, but the cross-correlation among the skew estimators increases with duration. Comparing the magnitude of $\rho$-*Correction* in Table 5.12 for the two data sets provides an idea of how sensitive the $\rho$-*Correction* term is to changes in the correlation structure of the data. While it is easy to see how $\rho$-*Correction* will behave in simple cases, it is not clear how realistic variance and correlation structures will affect that term.

Table 5.11 reports the empirical GLS ANOVA for the two example data sets. Here $SS_M(GLS) > 0$, indicating that the constant model is explaining some variability. This is clearly not true in a traditional sense. In Table 5.11 $SS_M(GLS) > 0$ because the constant model is a weighted mean, so that $SS_M(GLS)$ is the sum of squared differences between that weighted mean and the sample mean. In all cases in Table

5.11, the $CP(GLS)$ term is small, though generally noticeable. A key shortcoming of this and other empirical GLS ANOVAs is that $SS_E(GLS)$ does not make a distinction between model errors and sampling errors because it has no way to partition the $SS_E(GLS)$ between the two.

**Table 5.11:** Empirical GLS ANOVA for the 1-Day and 30-Day duration skews for the constant, linear, and non-linear elevation models.

| | 1-Day | | | 30-Day | | |
|---|---|---|---|---|---|---|
| Source | Constant | Linear Elevation | NL Elevation | Constant | Linear Elevation | NL Elevation |
| $SS_M(GLS)$ | 0.19 | 4.09 | 4.90 | 0.02 | 1.90 | 2.26 |
| $SS_E(GLS)$ | 7.64 | 3.19 | 2.94 | 5.75 | 3.59 | 3.45 |
| $CP(GLS)$ | -0.37 | -0.17 | -0.39 | -0.05 | 0.24 | 0.02 |
| $SS_T$ | 7.45 | 7.45 | 7.45 | 5.72 | 5.72 | 5.72 |

**Table 5.12:** B-GLS pseudo ANOVA for the 1-Day and 30-Day duration skews for the constant, linear and non-linear elevation models

| | 1-Day | | | 30-Day | | |
|---|---|---|---|---|---|---|
| Source | Constant | Linear Elevation | NL Elevation | Constant | Linear Elevation | NL Elevation |
| $SS(Model)$ | 0.00 | 2.43 | 3.10 | 0.00 | 0.74 | 1.03 |
| $SS(Model\ Error)$ | 3.65 | 1.22 | 0.55 | 1.54 | 0.80 | 0.51 |
| $SS(Sample\ Error)$ | 6.30 | 6.30 | 6.30 | 6.07 | 6.07 | 6.07 |
| $SS(Total)$ | 9.95 | 9.95 | 9.95 | 7.61 | 7.61 | 7.61 |

Table 5.12 reports the B-GLS pseudo ANOVA proposed by Gruber et al. [2007]. Unlike the empirical ANOVA, this analysis can and does distinguish between the model error and the sampling error. Also, this analysis correctly identifies that the constant model does not explain any variability by considering the modeled variance associated with each term. This analysis is somewhat divorced from the actual data

because it reports the variability that would be expected in a random sample with the variance structure of the estimated model.

One concern is that this analysis does not account for cross-correlation of the sampling errors [Gruber et al., 2007]. If sampling errors are positively correlated between sites, $SS(Sample\ Error)$ and $SS(Total)$ are likely to be too large. Thus this theoretical ANOVA will not match a corresponding empirical ANOVA.

In Table 5.12 $SS(Total)$ is not linked in any way to $SS_T$ or $E[SS_T]$; in certain cases $SS(Total)$ can be quite different than $SS_T$. It is not clear this is a problem, but it is an issue that should be understood. Much of the information provided by the B-GLS pseudo ANOVA does not depend on $SS(Total)$. For example, pseudo $R^2$ depends only on $SS(Model)$ and $SS(Model\ Error)$. Still a theoretical analysis with a strong tie to the traditional $SS_T$ is appealing.

**Table 5.13:** Alternative B-GLS pseudo ANOVA based on $E[SS_T]$ for the 1-Day and 30-Day duration skews for the constant, linear, and non-linear elevation models (after Table 5.9)

| | 1-Day | | | 30-Day | | |
|---|---|---|---|---|---|---|
| Source | Constant | Linear Elevation | NL Elevation | Constant | Linear Elevation | NL Elevation |
| $SS_*(Model)$ | 0.00 | 4.09 | 4.89 | 0.00 | 1.88 | 2.25 |
| $SS_*(Model\ Error)$ | 3.56 | 1.20 | 0.54 | 1.50 | 0.78 | 0.50 |
| $SS_*(Sample\ Error)$ | 6.17 | 6.17 | 6.17 | 5.94 | 5.94 | 5.94 |
| $\rho$-Correction | -1.78 | -1.78 | -1.78 | -2.09 | -2.09 | -2.09 |
| $SS_*(Total)$ | 7.95 | 9.68 | 9.82 | 5.35 | 6.51 | 6.60 |

Table 5.13reports the alternative pseudo ANOVA based on $E[SS_T]$ proposed in Section 5.2.3. Unlike, the B-GLS pseudo ANOVA proposed in Gruber et al. [2007], the analysis in Table 5.13 is based on the expectation of the various sum of squares that would be computed from the sample, including $E[SS_T]$. The terms which

describe model error and sampling error, $SS_*(Model\ Error)$ and $SS_*(Sample\ Error)$ respectively, are nearly identical to those in Table 5.12. However, there is now a correction term for the covariance of the sampling errors, *ρ-Correction*. This term should not be neglected; it is about 33% of $SS_*(Sample\ Error)$. In the table the *ρ-Correction* does not depend on the model, but it does increases with increasing duration. This occurs because the average cross-correlation of the observations also increases with duration. This is interesting because the average sampling errors are actually smaller for the longer duration, but the *ρ-Correction* still becomes larger.

The troubling aspect of Table 5.13 is that $SS_*(Total)$ depends on the selected model. This is clearly not reasonable: the total sum of squares should depend on the data, or the variance structure of the data, not the postulated model. This variation occurs because increases in $SS_*(Model)$ are not perfectly balanced by decreases in $SS_*(Model\ Error)$. The change in $SS_*(Total)$ is most dramatic when comparing the constant model to models with explanatory variables; but the fact that it changes at all is a concern. The discrepcency should not be surprising: $SS_*(Model\ Error)$ is based on the estimate of $\sigma_\delta^2(k)$ from a Bayesian analysis with an informative prior and $SS_*(Model)$ is based on an estimate of $\beta$. For small $n$ the prior of $\sigma_\delta^2(k)$ will have a large influence on $\hat{\sigma}_\delta^2(k)$, which would distort $SS_*(Model\ Error)$. For large $n$, small estimation errors in $\hat{\sigma}_\delta^2(k)$ can distort $SS_*(Model\ Error)$ when multiplied by $(n-1)$. Estimation errors in $b_J$ can have a great influence on $SS_*(Model)$. As a result $SS_*(Model)$ and $SS_*(Model\ Error)$ are not perfectly balanced.

The primary differences between the B-GLS pseudo ANOVA proposed by Gruber et al. [2007] and the alternative B-GLS pseudo ANOVA lays in the estimation of the variability explained by the model and the correction for correlation among the sampling errors. The alternative's estimate of the variability explained by the model, $SS_*(Model)$, causes $SS_*(Total)$ to vary depending on the postulated model. This is clearly not acceptable, and leads to our rejection of the alternative B-GLS pseudo ANOVA.

That said, the correction for correlation among the sampling errors could prove useful if added to the B-GLS pseudo ANOVA in Gruber et al. [2007]. It is true that it will not change the analyst's understanding of the breakdown of variability between sampling errors and model errors. Nor will it affect the judgment of whether an OLS analysis is sufficient. It will, however, inform the analyst of the impact of sampling error covariance, an issue which necessitated major methodological adaptations reported in this thesis. The current B-GLS pseudo ANOVA table does not reflect this potentially critical issue. Simply adding the correction term does not seem appropriate, as there are differences between $SS_*(Sample\ Error)$ and $SS(Sample\ Error)$, which could be substantial for small $n$. Instead, define a new ratio called the covariance loss factor (CLF):

$$CLF = \frac{\rho - Correction}{SS_*(Sample\ Error)} = \left(\frac{n-1}{n}\right)\left(\frac{\rho - Correction}{SS(Sample\ Error)}\right) \tag{5.51}$$

The CLF describes the fraction of adjusted sampling error variability ($SS_*(Sample\ Error)$) which has been lost (or gained) due to the correlation among the sampling errors. This can help indicate how much smaller the expected sum-of-

squared variability would be if we included the cross-correlation of the sampling errors in the ANOVA analysis. The second equality puts the covariance equality factor in terms of the Gruber et al. [2007] $SS(Sample\ Error)$. A CLF < 10% would suggest that the B-GLS pseudo ANOVA and the alternative would be very similar. When CLF is perhaps greater than 25%, one should recognize that the B-GLS pseudo ANOVA describes the relative average variation in a random observation due to the model, the model error, and due to sampling error – but it does not describe the expected sum of squared values for a set of observations with cross-correlated sampling errors. In general, if the average cross-correlations remain the same, CLF should increase with the number of sites n included in an analysis. With further experience it may be possible to understand how the magnitude of covariance loss factor should influence the choice of WLS, GLS or WLS/GLS analysis. Thus, Table 5.14 is a proposed modification of the B-GLS pseudo ANOVA proposed in Gruber et al. [2007].

Table 5.13 is identical to the B-GLS pseudo ANOVA proposed by Gruber et al. [2007] (Table 5.8), except that it includes the new covariance loss factor. By including $SS(Total)$ and $CLF$, Table 5.13 answers two potential questions: First, for a typical site, how much of the variability in the skew at that site is explained by the model, model error, and sampling error. The B-GLS pseudo ANOVA answers that question, and rather than being written in terms of variances and average variances, the B-GLS pseudo ANOVA is scaled to look like a standard ANOVA. The second question is how much total variability is expected in the data and how much of this is described by the model, model error, sampling error, and the covariance of sampling

errors. $CLF$ answers this question by indicating by how much the B-GLS pseudo ANOVA overestimates the expected sum of squares of a set of observations with correlated errors. Table 5.14 reports the proposed modification of the B-GLS pseudo ANOVA table for the numerical examples consider in this section.

**Table 5.14:** Proposed modification of the Gruber et al. [2007] pseudo ANOVA

| Source | Sum of Squares |
|---|---|
| $SS(Model)$ | $n\left[\hat{\sigma}_\delta^2(0) - \hat{\sigma}_\delta^2(k)\right]$ |
| $SS(Model\ Error)$ | $n\hat{\sigma}_\delta^2(k)$ |
| $SS(Sample\ Error)$ | $\sum_{i=1}^{n} var(\tilde{y}_i)$ |
| $SS(Total)$ | Sum of Above |
| $CLF$ | $\left(\frac{n-1}{n}\right)\left(\frac{\rho - Correction}{SS(Sample\ Error)}\right)$ |

The top half of Table 5.14 is identical to the B-GLS pseudo ANOVA, (Table 5.11). The new line for CLF indicates that because of the cross-correlation, we expect sampling errors variability to decrease about 27-28%. A nice feature of the CLF is that it can be computed before any analysis is started, because it only depends on the sampling error covariance matrix.

**Table 5.15:** Numerical application of the proposed modification of the B-GLS pseudo ANOVA

| | 1-Day | | | 30-Day | | |
|---|---|---|---|---|---|---|
| Source | Constant | Linear Elevation | NL Elevation | Constant | Linear Elevation | NL Elevation |
| $SS(Model)$ | 0.00 | 2.43 | 3.10 | 0.00 | 0.74 | 1.03 |
| $SS(Model\ Error)$ | 3.65 | 1.22 | 0.55 | 1.54 | 0.80 | 0.51 |
| $SS(Sample\ Error)$ | 6.30 | 6.30 | 6.30 | 6.07 | 6.07 | 6.07 |
| $SS(Total)$ | 9.95 | 9.95 | 9.95 | 7.61 | 7.61 | 7.61 |
| $CLF$ | -0.28 | -0.28 | -0.28 | -0.27 | -0.27 | -0.27 |

### *Section 5.2.5 Reflection of B-GLS ANOVA Tables*

Previous sections of this chapter have provided both theoretical insight and numerical experience related to reasonable definitions of a ANOVA for the Stedinger–Tasker GLS regression framework. A fundamental problem is that the sampling errors are not observed, and thus an empirical ANOVA such as that typically generated for OLS analyses cannot be generated. A reoccurring issue is what one hopes the ANOVA table is intended to illustrate. B-GLS pseudo ANOVA in Table 5.7 is very attractive because it is relative simple and highlights the key sources of variability in the data: variation explained by the model, variability attributed to model errors, and the expected variability in the sampling errors. For all models the sum of model variability and model error is represented by $n\hat{\sigma}_\delta^2(0)$, which is the best B-GLS estimate of the variability in the $\tilde{y}_i$ values obtained by a B-GLS analysis of the constant model. For other regression models, model variability and model error variability are divided as $n\left[\hat{\sigma}_\delta^2(0) - \hat{\sigma}_\delta^2(k)\right]$ and $n\hat{\sigma}_\delta^2(k)$, respectively, where $\hat{\sigma}_\delta^2(k)$ is the best B-GLS estimate of the model error variance for the model with *k* parameters. To the sum of model and model error variability is added the expected sum of squared sampling errors,

$$E\{\sum_{i=1}^{n} \eta_i^2\} \;=\; \sum_{i=1}^{n} var(\tilde{y}_i)$$

where $E\{\eta_i^2\} = var(\eta_i)$ has been denoted as $var(\tilde{y}_i)$ throughout this thesis.

A concern had been that these three terms ignored the cross-correlation among the sampling errors, and how those error interact with estimated parameters. Here Table 5.8 is particularly informative. If we try to match the Table 5.7 theoretical sums of squares (described as *n* times the variance of each variable), to the expected value

of the empirical sums of squares in Table 5.9, we see that its use of the sample estimate of the average of the $\tilde{y}_i$ that causes the differences. Use of the average results in the (*n-1*) factors in Table 5.8 that appears in the $SS_*(Model\ Error)$ and $SS_*(Sample\ Error)$ rows. The (*n-1*) factor mimics the loss of one degree of freedom that is a consequence of computing the sum of squares about the sample mean, and is a relatively small correction. The big impact results from the use of the sample mean when envisioning the sum of squares of the sampling errors. Were the n sampling errors $\eta_\iota$ known, then the expected sum of their squares about their theoretical mean of zero would be

$$E\{\sum_{i=1}^{n}\eta_i^2\} = \sum_{i=1}^{n} var(\tilde{y}_i)$$

However, when we subtract the sample average of these errors from all the values, we obtain

$$E\{\sum_{i=1}^{n}(\eta_i-\bar{\eta})^2\} = E\{\sum_{i=1}^{n}\eta_i^2 - n\bar{\eta}^2\}$$

$$= \sum_{i=1}^{n} var(\tilde{y}_i) - \frac{1}{n}\{\sum_{i=1}^{n} var(\tilde{y}_i) + \sum_{i=1}^{n}\sum_{j\neq i}\mathbf{\Lambda}_{GLS}(i,j)\}$$

$$= (n-1)\sum_{i=1}^{n}\frac{var(\tilde{y}_i)}{n} - \frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}\mathbf{\Lambda}_{GLS}(i,j)$$

which is simply $SS_*(Sample\ Error)$ plus the *$\rho$-correction* in Table 5.8. Thus the B-GLS pseudo ANOVA proposed by Gruber et al. [2007] correctly describes the theoretical variability due to the model, the model errors, and the sampling errors if one imagines centering the errors about their true mean of zero. However, its estimate of the total sum of squares is too large if one envisions centering the sum of squares about a sample average value of $\tilde{y}_i$ because the correlations among all the sampling errors would reduce the computed sum of squares. Thus it is reasonable to argue that

the B-GLS pseudo ANOVA proposed by Gruber et al. [2007] is the appropriate representation of the total variability in ones data, which would on average be underestimated by the empirical sum of squares. However, the distortion in the expected empirical sum of squares is a useful measure of the impact of cross-correlation on the analysis and has been included in the misrepresentation of beta-variance statistics discussed in Chapter 3.

***Section 5.2.6: Conclusions***

When analyzing the California skew Data described in Chapter 4, it was observed that the B-GLS pseudo ANOVA table was different than the empirical ANOVA table. While this is not in itself problematic, it is useful to understand why the two tables differ. This section explores the issue of ANOVA as a GLS regression diagnostic for a range of GLS applications. ANOVA are divided into two types: empirical and theoretical. Empirical approaches are based solely on the data, whereas theoretical approaches are based on the hypothesized error variance structure with the estimated parameters. Thus empirical ANOVA report the variability in a sample, while theoretical ANOVA report the expected variability in a random sample with a given error variance structure.

Theoretical ANOVA are necessary in the Stedigner and Tasker [1985] regression framework because sampling error and model error variability are not easily segregated with empirical methods. It was noted that the theoretical ANOVA proposed by Gruber et al. [2007] (denoted B-GLS pseudo ANOVA) did not include all of the terms in $E[SS_T]$; in particular it omits $\rho - Correction.$ A new theoretical ANOVA based on an approximation of $E[SS_T]$ was derived, denoted alternative B-

GLS pseudo ANOVA. This new alternative is ultimately rejected because the estimate of the total variability in the data is highly dependent on the hypothesized model. However, a new covariance loss factor, which estimates the fraction of sampling error variability lost due to positive correlation among the sampling errors has been added to the B-GLS pseudo ANOVA table. The new factor provides the analyst with an estimate of the factor by which the sum of squares of the sampling errors in the data set is decreased relative to the value that would be expected if sampling errors were independent.

### *Section 5.3: Consistency of the Regional Skew Models Across Durations*

This section examines consistency across durations of the regional skew models developed in Chapter 4. This was a concern raised during the review process for Lamontagne et al. [2012], which reports most of the analyses in Chapter 4. In particular, it addresses the concern that the skew models are not ordered by duration as one might intuitively expect. For example one might expect that as duration increased the magnitude of skew models might either increase or decrease monotonically. The underlying concern is that subsequent flood frequency analyses for each duration is consistent with each other, i.e. the $p^{th}$ annual exceedance probability (AEP) flood should decrease with increasing duration. The $p^{th}$ AEP is the flood which has probability $p$ of being exceeded in any given year. Thus the 0.01 AEP flood will be exceeded in any year with probability 1/100. This is also commonly called the 100 year flood.

Here the consistency concern is addressed in three ways: Section 5.3.1 considers the real-space characteristics of log-space models developed in Chapter 4.

Section 5.3.2 compares the estimated 0.01 AEP floods computed with the sample standard deviation and regional skew for a variety of study sites. Finally, Section 5.3.2 examines the effect of the log-space skew and log-space standard deviation on the relative magnitude of floods of different durations.

Before entering a discussion of regional skew model trends across durations, it is important to recall that the parameters of the final regional models from each duration were not statistically different from each other, meaning that the simple confidence intervals for the parameters overlapped. This is not evidence that a single model should necessarily be used for all durations: the model parameters are not significantly different from many alternative values. However, it is a caution not to attribute too much to differences between regional skew models for different durations.

### *Section 5.3.1: Real-space Characteristics of Regional log-space Skew Models across Study Durations*

This section examines the real-space characteristics of the log-space skew models developed in Chapter 4 for a variety of study basins across each of the five rainfall flood durations. This is done by considering the real-space skew coefficient of a log-Pearson Type III with a log-space skew coefficient equal to the regional skew from Chapter 4.

The real-space skew of the log-Pearson Type III distribution depends on both the log-space skew coefficient and the log-space standard deviation [Griffis and Stedinger 2007; Stedinger et al., 1993]. One should not draw conclusions about the shape of the fitted distributions for the various durations by simply comparing the log-space skew coefficients.

To help clarify this issue, the real-space skew coefficients associated with the log-space regional skew were calculated for each of the five study durations at twelve representative sites, which had their own at-site log-space standard deviations. Sites were selected that had long record lengths and spanned a wide range of mean basin elevations. Recall that mean basin elevation is an explanatory variable in the final models selected in Chapter 4. Those models essentially assign a constant skew for low elevations and a constant skew for high elevations, with a rapid transition zone between the two. Four sites were selected for each of the three elevation types: low, transition, and high. Characteristics of the sites are summarized in Table 5.16.

**Table 5.16:** Summary of Twelve Representative Basins from three elevation categories.

| Site Number | Site Name | Record Length | Mean Basin Elevation (ft) | Category |
|---|---|---|---|---|
| 15 | Bear R. near Wheatland | 103 | 2250 | Low |
| 30 | Calaveras R. at Hogan Dam | 96 | 1991 | Low |
| 43 | Cache Ck. at Clear Lake | 87 | 2004 | Low |
| 50 | Arroyo Seco R. near Soledad | 107 | 2494 | Low |
| 12 | Butte Ck. near Chico | 78 | 3717 | Trans. |
| 33 | Fresno R. near Knowles | 76 | 3201 | Trans. |
| 10 | Big Chico Ck. near Chico | 77 | 3111 | Trans. |
| 45 | M Fork Eel R. near Dos Rios | 43 | 3685 | Trans. |
| 13 | Feather R. at Oroville Dam | 107 | 5031 | High |
| 24 | Merced R. at Exchequer Dam | 107 | 5473 | High |
| 25 | Tuolumne R. at Don Pedro Dam | 112 | 5882 | High |
| 31 | Mokelumne R. at Camanche Dam | 104 | 4918 | High |

The log-space mean and standard deviation was computed for each site and each duration using the Expected Moments Algorithm (EMA), following the same censoring recommendations used in Chapter 4. The real-space skew coefficient is then computed using the log-space sample standard deviation and the log-space regional skew coefficient. Figure 5.5 plots the log-space skew coefficients for low,

transition, and high elevation basins for each of the five study durations. Figure 5.6, Figure 5.7, and Figure 5.8 plot the real-space regional skew coefficient for the low, transition, and high elevations respectively.

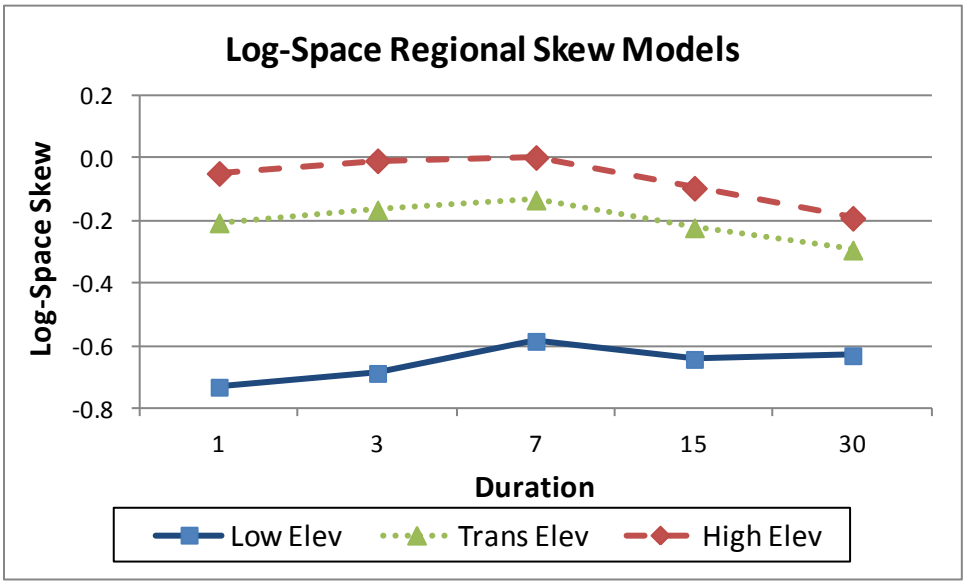


**Figure 5.5:** Log-Space Regional Skew Coefficient for Low, Transitional, and High Elevation Basins.

In log-space, the regional skew coefficients increase with duration from 1-Day to 7-Day, then generally decrease with duration from 7-Day to 30-Day. The exception is low elevation 30-Day skew, which is greater than low-elevation 15-Day. The precision of the model parameters makes it difficult to determine if this is a real difference or a result of sampling error in the computed parameters.

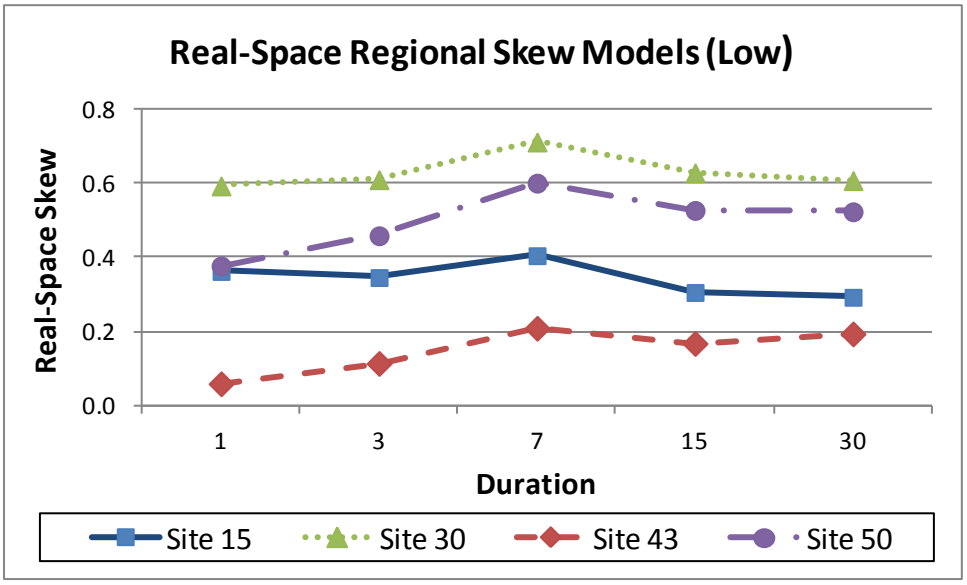


**Figure 5.6:** Real-Space Regional Skew Coefficient for four Low Elevation Basins, using regional log-space skew and sample log-space standard deviation.

For low-elevation sites, the real-space skew coefficient remains relatively constant across durations. In all cases the 7-Day skew coefficient is largest, and there is almost no difference between the 15-Day and 30-Day regional skew coefficient.

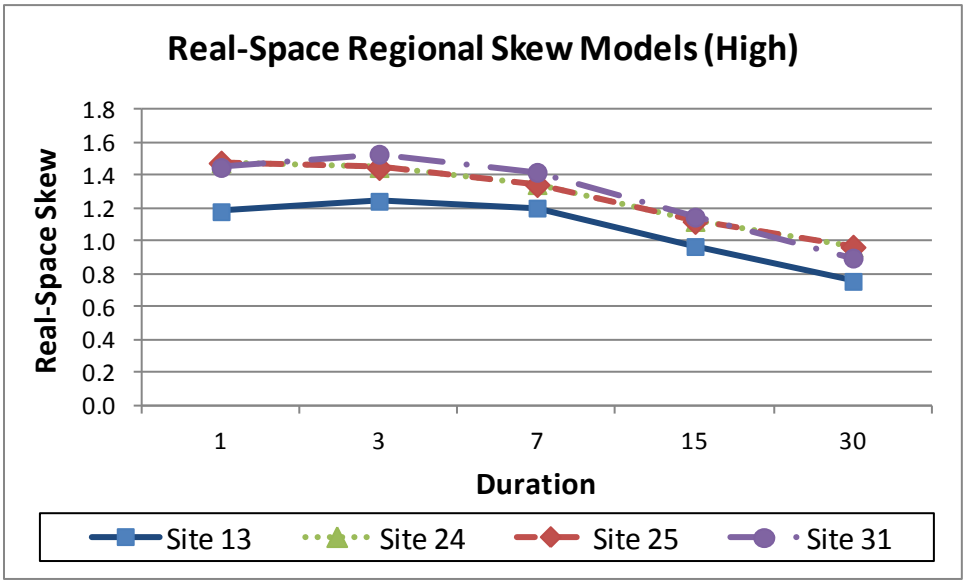


**Figure 5.7:** Real-Space Regional Skew Coefficient for four High Elevation Basins, using regional log-space skew and sample log-space standard deviation.

High elevation real-space regional skew is generally greatest for the 3-Day duration, before decreasing with increasing duration. Real-space regional skew values change much more with duration for high elevation sites than for low-elevation sites. This was also the case in log-space (see Figure 5.5). This indicates that the shape of the distribution of rainfall floods is more uniform across durations for low elevation sites than high elevation sites. The hydrologic reasons for this are difficult to determine, but elevation in California often dictates the degree to which rain-snow interactions impact the flood hydrology [Parrett et al., 2011; Lamontagne et al., 2012].

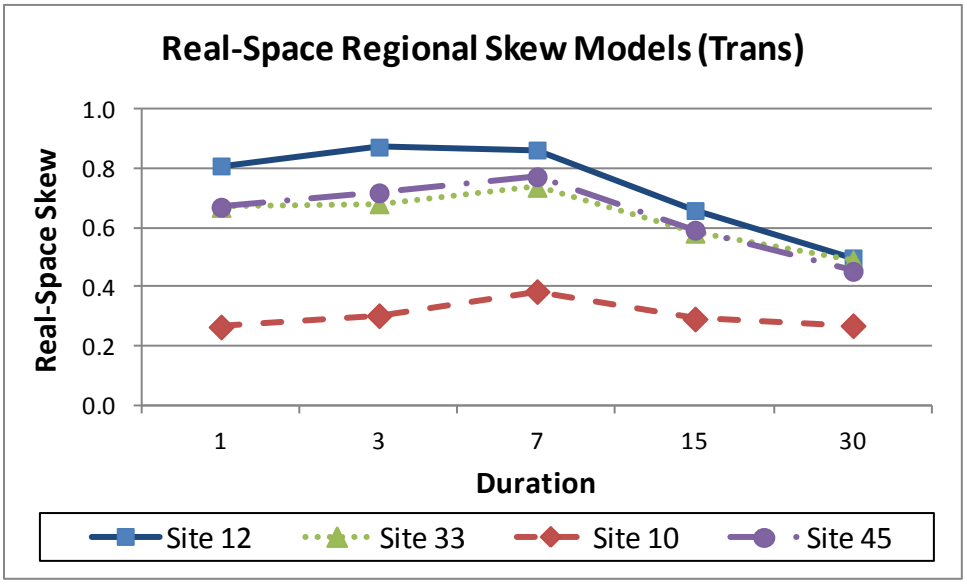


**Figure 5.8:** Real-space Regional Skew Coefficient for four Transitional Elevation Basins, using regional log-space skew and sample log-space standard deviation.

As one would expect, the trend in real-space regional skew coefficient across durations for the transitional sites is a mixture of the trends observed for high and low sites. Site 12 has the highest mean basin elevation of all of the transitional basins (3717 ft). Not surprisingly, the trend in real-space regional skew coefficient for Site 12 is very similar to that for high basins, i.e. 3-Day duration has the maximum real-space skew with decreasing skew for longer durations. Site 11 has the lowest mean

basin elevation of all of the transitional basins (3111 ft). The real-space regional skew values for Site 11, i.e. more uniform real-space skew, with 7-Day having the maximum magnitude. This is similar to what is observed in Figure 5.6.

Overall, the trend in real-space skew across durations is not erratic for any of the representative sites. One would be very concerned if the regional skew for different durations at the same site were drastically different, but they are not. The trend in both real-space and log-space regional skew values are not monotonic with duration, but there is no obvious or compelling reason why it should be. The underlying hydrologic reasons for the observed trends in regional skew are not clear, though elevation seems to affect the spread of long and short duration regional skews. It is speculated this might because of the impact of snowmelt hydrology on distribution of floods of different durations at high elevations, but it is difficult to draw hydrologic meaning from the skew coefficient of annual maximum flows.

***Section 5.3.2: Comparison of 100-year flood estimates for different durations computed with regional skew model***

The annual maximum d-Day rainfall flood record is generated by averaging rainfall flood records over a sliding d-Day window, then selecting the maximum for each water year. Thus, the magnitude of the d-Day rainfall flood cannot increase with duration:

$$Q_i(t) \geq Q_j(t) \qquad \text{if } i < j \tag{5.52}$$

Where $i$ and $j$ are rainfall flood durations, and $Q_d(t)$ is the maximum d-Day rainfall flood (flow rate) for year $t$ and duration. For this reason, flood quantiles for a fixed AEP, $p$, should not increase with decreasing duration:

$$Q_{1-p}(i) \geq Q_{1-p}(j) \qquad \text{if } i < j \tag{5.53}$$

Where $Q_{1-p}(d)$ is the $p$ AEP flood for duration d. A concern is that the flood quantiles computed using the regional skew models in Chapter 4 do not provide $Q_{1-p}$ which violate Equation (5.53).

As a check, the 0.01 AEP flood (i.e. 100-year flood) is computed using the regional skew for the 12 representative sites and the five study durations. This involves using the at-site sample mean, at-site sample standard deviation, and the regional skew for each site and duration to estimate the 0.01 AEP flood. This is not the procedure recommended by Bulletin 17B, because the at-site sample skew is not used in the computation. This example is only intended to see if the regional skew itself causes inconsistencies. Figure 5.9, Figure 5.10, and Figure 5.11 plot the 0.01 AEP flood for each of the study durations for low, high, and transitional representative basins respectively. Note that for every basin, the magnitude of the 0.01 AEP flood decreases with increasing duration.

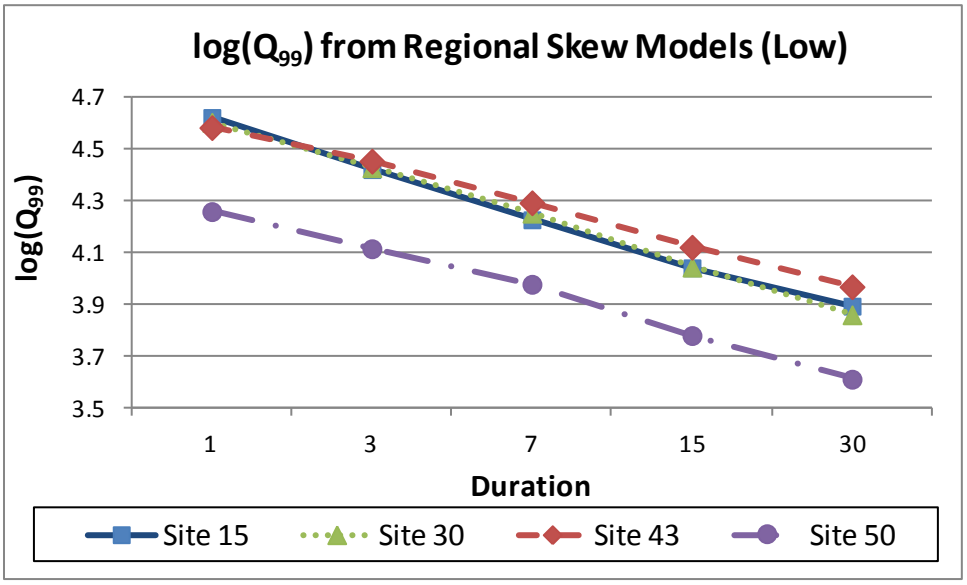


**Figure 5.9:** Rainfall Duration Flood 0.01 AEP for four low elevation sites, computed using sample log-space mean and standard deviation, and regional log-space skew coefficient.

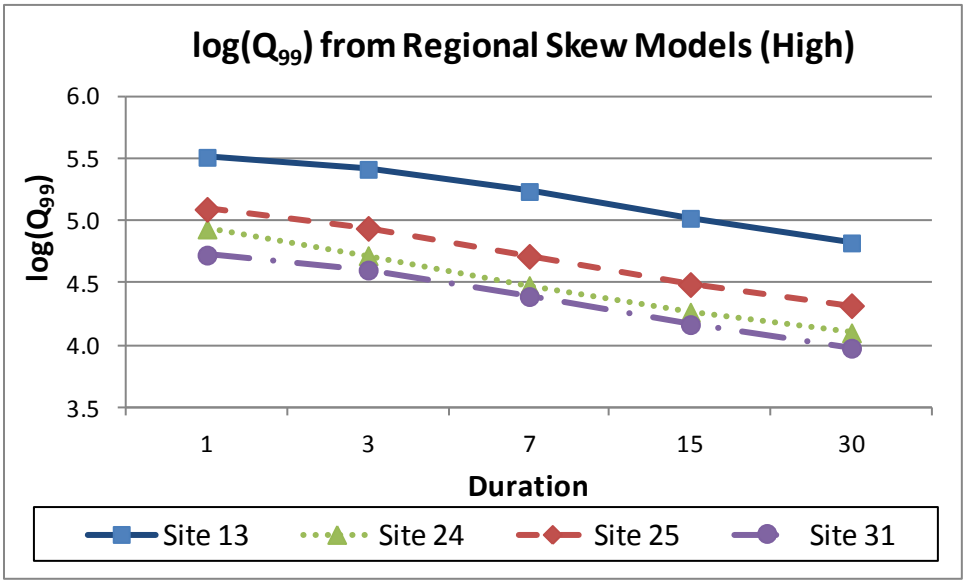


**Figure 5.10:** Rainfall Duration Flood 0.01 AEP for four high elevation sites, computed using sample log-space mean and standard deviation, and regional log-space skew coefficient.

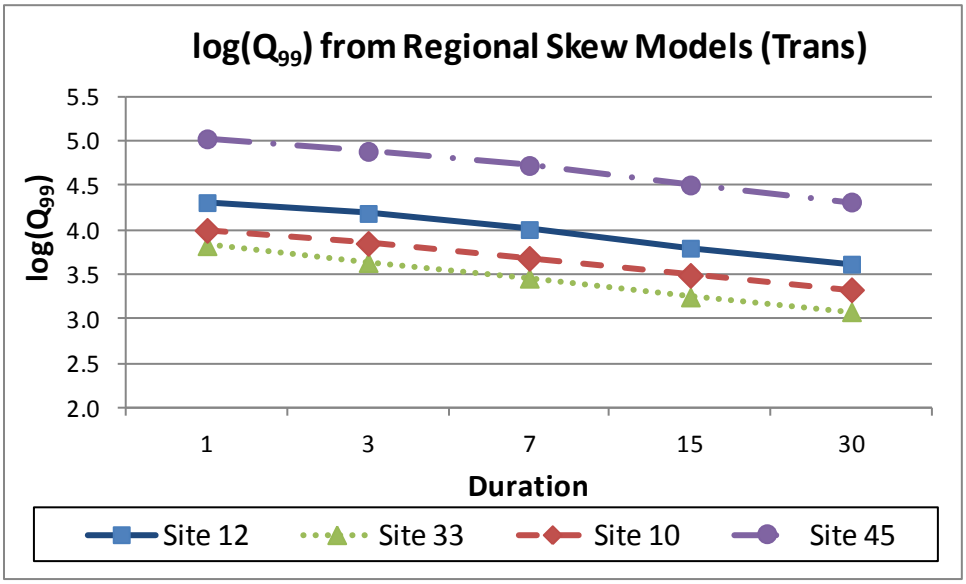


**Figure 5.11:** Rainfall Duration Flood 0.01 AEP for four transition elevation sites, computed using sample log-space mean and standard deviation, and regional log-space skew coefficient.

It is encouraging that the magnitude of 100-year flood decreases monotonically with duration in every case. Thus, it seems relatively certain that the trend of regional skew in duration is not causing inconsistencies in estimation of the 0.01 AEP. Use of the sample skew might cause such inconsistencies, but that is a separate issue from the regional skew, which is the concern here.

***Section 5.3.3: Comparison of p-year flood estimates for different durations computed with regional skew model for three representative sites***

Many applications often require estimation of flood quantiles more extreme than the 0.01 AEP flood [Calzascia and Fitzpatrick, 1989]. This section will explore the consistency of various AEP flood estimators computed with the regional skew for three representative sites: Site 50, Site 12, and Site 25. These sites represent low, transition, and high elevation sites respectively, and all have long record lengths.

Following the same procedure described in Section 5.3.2, the 0.02, 0.01, 0.004, 0.002, and 0.001 AEP floods are computed for each site and duration. Those are the

50-, 100-, 250-, 500-, and 1000-year floods. Figure 5.12, Figure 5.13, and Figure 5.14 plot the computed AEP for Site 50 (low), Site 12(transitional), and Site 25(high) respectively.

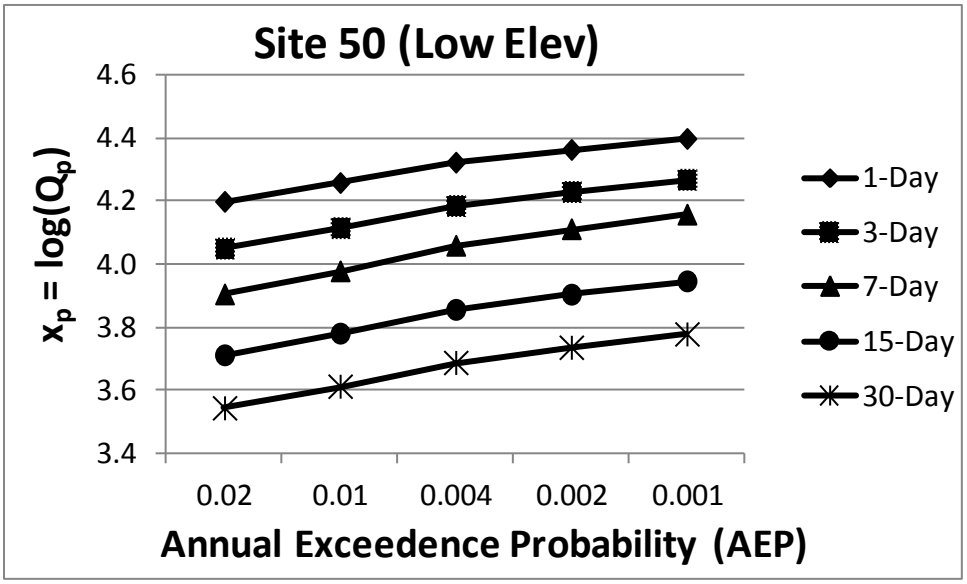


**Figure 5.12:** Site 50 AEP flood quantiles for five durations, computed using sample log-space mean and standard deviation and regional log-space skew coefficient.

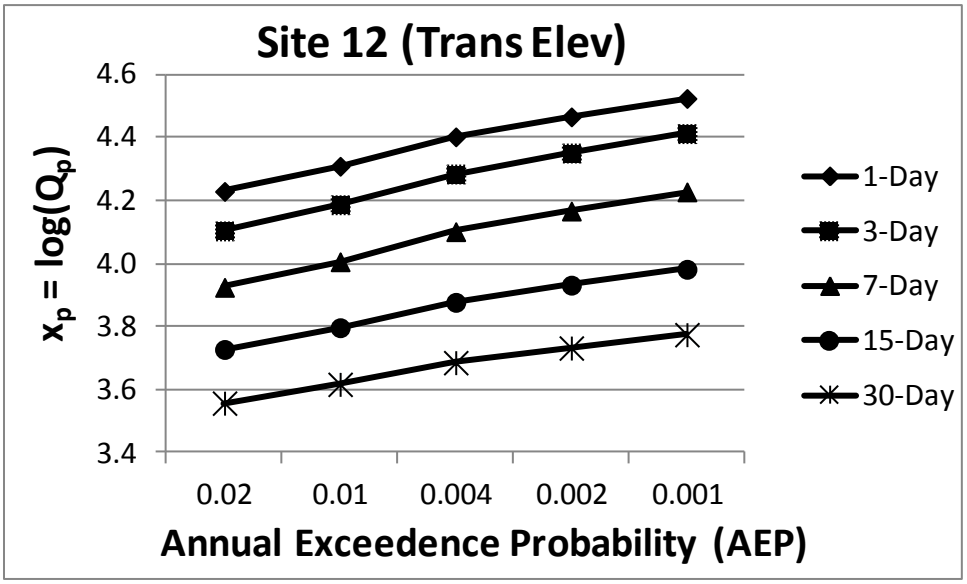


**Figure 5.13:** Site 12 AEP flood quantiles for five durations, computed using sample log-space mean and standard deviation and regional log-space skew coefficient.

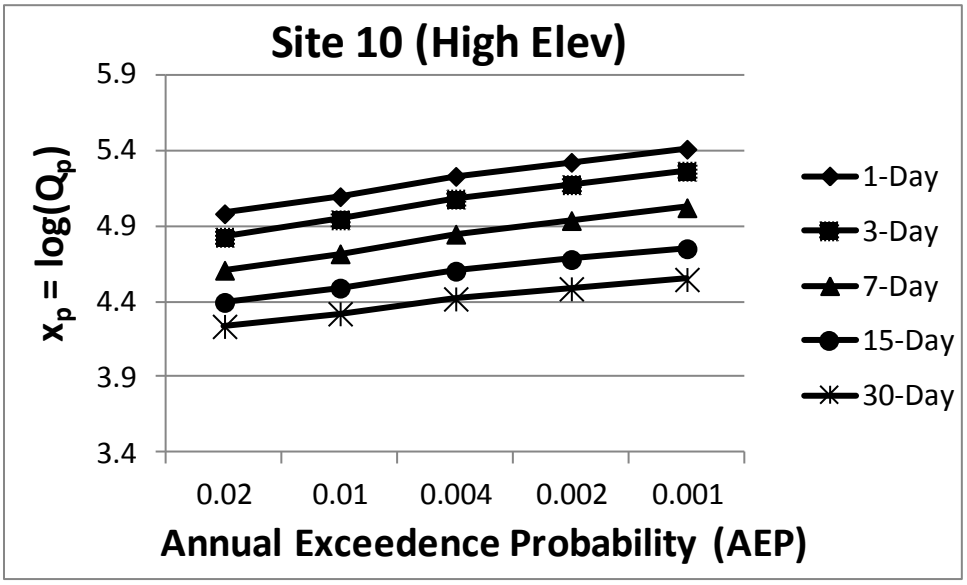


**Figure 5.14:** Site 25 AEP flood quantiles for five durations, computed using sample log-space mean and standard deviation and regional log-space skew coefficient.

For Equation (5.53) to hold, no two lines in Figure 5.12-Figure 5.14 should touch. It is encouraging that they have not, meaning that out to the 0.001 AEP flood (1000-year flood), the regional skew has not caused the feared inconsistencies. Again, this has not replicated the Bulletin 17B procedure, which would involve use of the sample skew coefficient. Instead, this analysis is concerned only with potential inconsistencies caused by the non-monotonic trends in regional skew magnitude. It is encouraging that in the three cases tested in this section, Equation (5.53) holds and the feared inconsistencies were not observed out to the 0.001 AEP flood (1,000-year flood).

***Conclusion***

This chapter addresses several issues which were raised during the review process for Lamontagne et al. [2012]. Section 5.1 considers whether the remarkably high ERL and low $VP_{new}$ reported in Chapter 4 are reasonable. The relationship

between ERL, $VP_{new}$, and the magnitude of the skew model is explored. Also, a simple approximation of the final model in Chapter 4 is proposed. This model returns an approximate $VP_{new}$ which is very similar to the values reported in Chapter 4, confirming that the low $VP_{new}$ is reasonable. Section 5.2 reexamines the Pseudo ANOVA table, compares it to a traditional ANOVA table, and an addition term to account for the cross-correlation of sampling errors. Finally, Section 5.3 examines whether the trend of the regional skew models across flood durations is reasonable, and whether they lead to inconsistencies in subsequent flood frequency analyses. The real-space regional skew values for various representative sites are compared, and flood quantiles are computed using the regional skew coefficient. Inconsistencies are not observed, confirming that while trends across durations are unexpected and hard to explain, they do not seem to return inconsistent flood frequency results.

# REFERENCES


Baxter, N.D. and J.G. Cragg. 1970. Corporate Choice Among Long-Term Financing Instruments. The Review of Economics and Statistics, 52(3), 225-235.

Blomquist, N.S. 1980. A Note on the Use of the Coefficient of Determination. Scandinavian J. of Econ. 82(2), 409-412.

Bretscher, O. (2008), *Linear Algebra with Applications*. Pearson, New York, NY.

Buse, A. (1973). Goodness of Fit in Generalized Least Squares Estimation. The American Statistician, 27(3), 106-108.

Calzascia, E.R., Fitzpatrick, J.A. (1989). "Hydrologic Analysis within California's Dam Safety Program", ASDSO Western Regional Conference and Dam Safety Workshop, May 1-3, 1989, Sacramento, CA.

Cameron, A.C. and Trivedi, P.K. (1998). *Regression Analysis of Count Data*. Cambridge University Press, Cambridge, UK.

Dobson, A.J. (2002). *An Introduction to Generalized Linear Models*. Chapman & Hall, Boca Raton, FL.

Draper, N.R. and Smith, H. (1967). *Applied Regression Analysis*. John Wiley & Sons, Inc., New York, N.Y.

Gelman A. and I. Pardoe. (2006). Bayesian Measures of Explained Variance and Pooling in Multilevel (Hierarchical) Models. Technometrics. 48(2). 241-251.

Greene, W.H. (2008). *Econometric Analysis*. Pearson-Prentice Hall, Upper Saddle River, N.J.

Griffis, V. W., and Stedinger, J. R. (2007). "Log-Pearson Type 3 Distribution and Its Application in Flood Frequency Analysis. I: Distribution characteristics." *J. Hydrol. Engineering*., 12 (5), 482–491.

Griffis, V.W., and J. R. Stedinger, (2009), The Log-Pearson Type 3 Distribution and its Application in Flood Frequency Analysis, 3. Sample Skew and Weighted Skew Estimators, *J. of Hydrol. Engineering* 14(2), pp. 121-130.

Gruber, A.M., D.S. Reis Jr., and J. R. Stedinger (2007), Models of regional skew based on Bayesian GLS regression, World Environmental & Water Resources Conference-Restoring out Natural Habitat, edited by K.C. Kabbes, Tampa, Florida May 15-18, Paper 40927-3285

Herr, D.G. (1986). On the History of ANOVA in Unbalanced, Factorial Designs: The First 30 Years. The American Statistician. 40(4). 265-270.

Interagency Advisory Committee on Water Data, 1982, Guidelines for determining flood-flow frequency, Bulletin #17B of the Hydrology Subcommittee, Office of Water Data Coordination: U.S. Geological Survey, Reston Virginia, 183 p. Available at *http://water.usgs.gov/osw/bulletin17b/dl_flow.pdf*

Jarrett, J.P. (1974). The Coefficient of Determination-Some Limitations. The American Statistician. 28(1), 19-20.

La Du, T.J. and J.S. Tanaka, Influence of Sample Size, Estimation Method, and Model Specification on Goodness-of-Fit Assessments in Structural Equation Models. Journal of Applied Psychology. 74(4). 625-635.

Lamontagne, J.R., J.R. Stedinger, C. Berenbrock, A.G. Veilleux, J.C. Ferris, and D.L. Knifong, 2012. Development of Regional Skews for Selected Flood Durations

for the Central Valley Region, California Based on Data Through Water Year 2008, U.S. Geological Survey Scientific Investigations Report 2012-5130 60 p.
Langsrud, Oyvind. (2003). ANOVA for unbalanced data: Use Type II instead of Type III sums of squares. Statistics and Computing 13. 163-167.
Long, J.S. (1997). *Regression Models for Categorical and Limited Dependent Variables*. Sage Publications, London, UK.
Liu, H., R.A. Davidson, D.V. Rosowsky, J.R. Stedinger. (2005). Negative Binomial Regression of Electric Power Outages in Hurricanes, Journal of Infrastructure Systems, 11(4), 258-267.
Menard, S. (2000). Coefficients of Determination for Multiple Logistic Regression Analysis, The American Statistician, 54(1), 17-24.
Parrett, C., A. Vellieux, , J. R. Stedinger, N. A. Barth, D. Knifong, , and J.C. Ferris, 2010. Regional Skew for California and Flood Frequency for Selected Sites in the Sacramento-San Joaquin River Basin Based on Data through Water Year 2006, OFR 2010, U.S. Geological Survey.
Reis Jr., D.S. 2005. "Flood Frequency Analysis Employing Bayesian Regional Regression and Imperfect Historical Information." PhD Dissertation, School of Civil and Environmental Engineering, Cornell Univ., Ithaca, N.Y.
Stedinger, J. R., 1983, Estimating a Regional Flood Frequency Distribution: Water Resources Research, v. 19, no. 2, p. 503-510.
Stedinger, J. R., and Tasker, G. D., 1985, Regional hydrologic analysis, 1, ordinary, weighted and generalized least squares compared: Water Resources Research ,v. 21, no. 9, p. 1421-1432. [with correction, Water Resources Research, v. 22, no. 5, p. 844, 1986.]
Veilleux, A. G. 2009. "Bayesian GLS Regression for Regionalization of Hydrologic Statistics, Floods and Bulletin 17 Skew." MS thesis, School of Civil and Environmental Engineering, Cornell Univ., Ithaca, N.Y.
Yates, F. (1934). The analysis of multiple classifications with unequal numbers in the different classes. Journal of the American Statistical Association. 29, 51-66.

***Chapter 5 Appendix***

This appendix derives $E[SS_T]$ for the hybrid WLS/GLS analysis used in Chapter 4, and the general WLS or GLS case. Start by recalling some definitions. Let $\widetilde{\boldsymbol{Y}}$ be a $n \times 1$ vector of observations. Assuming a linear model,

$$\widetilde{\boldsymbol{Y}} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \tag{5.54}$$

where $\boldsymbol{X}$ is a $n \times k$ matrix of basin characteristics, and $\boldsymbol{\beta}$ is a $k \times 1$ vector of model parameters, and $\boldsymbol{\varepsilon}$ is a $n \times 1$ vector of errors. Note that

$$\begin{aligned} E[\boldsymbol{\varepsilon}] &= \mathbf{0} \\ E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T] &= \boldsymbol{\Lambda} \end{aligned} \tag{5.55}$$

Let $\boldsymbol{\Lambda}_J$ be the covariance matrix of $\boldsymbol{\varepsilon}$ from a $J$-case analysis. Let $\boldsymbol{b}_J$ be the least-squares estimate of $\boldsymbol{\beta}$. Recall from Chapter 3,

$$\boldsymbol{b}_J = \left(\boldsymbol{X}^T\boldsymbol{\Lambda}_J^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{\Lambda}_J^{-1}\widetilde{\boldsymbol{Y}} = \boldsymbol{W}_J\widetilde{\boldsymbol{Y}} \tag{5.56}$$

Where $\boldsymbol{W}_J$ is the $J$-case weight matrix, defined as:

$$\boldsymbol{W}_J = \left(\boldsymbol{X}^T\boldsymbol{\Lambda}_J^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{\Lambda}_J^{-1} \tag{5.57}$$

The fitted model for the $J$-case is given by:

$$\widehat{\boldsymbol{Y}}_J = \boldsymbol{X}\boldsymbol{b}_J = \boldsymbol{X}\boldsymbol{W}_J\widetilde{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{W}_J(\boldsymbol{X\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{X\beta} + \boldsymbol{X}\boldsymbol{W}_J\,\boldsymbol{\varepsilon} \tag{5.58}$$

Let $\overline{\boldsymbol{Y}}$ be a $n \times 1$ vector containing the sample mean in each element. $\overline{\boldsymbol{Y}}$ is computed as

$$\overline{\boldsymbol{Y}} = \boldsymbol{N}\widetilde{\boldsymbol{Y}} = \boldsymbol{N}(\boldsymbol{X\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{NX\beta} + \boldsymbol{N\varepsilon} \tag{5.59}$$

where $\boldsymbol{N}$ is an $n \times n$ matrix containing $1/n$ in each element.

Suppose a WLS/GLS hybrid analysis, such as the procedure described in Chapter 4, is used. This means that a WLS analysis is used to estimate the model parameters, and a GLS analysis is used to estimate their precision. Generalizing this, suppose the $K$-case is used to estimate the model parameters and the $J$-case is used to estimate the precision of the model. This means that the fitted model is computed as

$$\widehat{\boldsymbol{Y}}_K = \boldsymbol{X\beta} + \boldsymbol{XW}_K\,\boldsymbol{\varepsilon} \tag{5.60}$$

where $E[\boldsymbol{\varepsilon\varepsilon}^T] = \boldsymbol{\Lambda}_J$.

Note that

$$E[SS_T] = E[SS_E(K,J)] + E[SS_M(K,J)] + 2E[CP(K,J)] \tag{5.61}$$

where $SS_E(K,J)$, $SS_M(K,J)$, and $CP(K,J)$ are the sum of squares due to errors, sum of squares due to the model, and the cross-product term for a hybrid $K/J$ case.

To estimate $E[SS_T]$, an approximation for each of the expressions on the RHS of ( will be derived. First consider $E[SS_E(K,J)]$. Recall that $SS_E(K,J) = \left(\widetilde{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_K\right)^T(\widetilde{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_K)$. Expanding the definition of $SS_E(K,J)$:

$$\begin{aligned} SS_E(K,J) &= \left(\widetilde{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_K\right)^T\left(\widetilde{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_K\right) \\ &= (\boldsymbol{X\beta} + \boldsymbol{\varepsilon} - \boldsymbol{X\beta} - \boldsymbol{XW}_K\,\boldsymbol{\varepsilon})^T(\boldsymbol{X\beta} + \boldsymbol{\varepsilon} - \boldsymbol{X\beta} - \boldsymbol{XW}_K\,\boldsymbol{\varepsilon}) \\ &= (\boldsymbol{\varepsilon} - \boldsymbol{XW}_K\,\boldsymbol{\varepsilon})^T(\boldsymbol{\varepsilon} - \boldsymbol{XW}_K\,\boldsymbol{\varepsilon}) \\ &= \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}^T\boldsymbol{XW}_K\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}^T\boldsymbol{W}_K^T\boldsymbol{X}^T\boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}^T\boldsymbol{W}_K^T\boldsymbol{X}^T\boldsymbol{XW}_K\boldsymbol{\varepsilon} \end{aligned} \tag{5.62}$$

Recall the property that the trace of a matrix is invariant under cyclic permutations (i.e. $trace(ABC) = trace(BCA)$) [Bretscher, 2009], and that the trace of a scalar is the scalar:

$$\begin{aligned} SS_E(K,J) &= trace(\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}) - trace(\boldsymbol{\varepsilon}^T\boldsymbol{X}\boldsymbol{W}_K\boldsymbol{\varepsilon}) - trace(\boldsymbol{\varepsilon}^T\boldsymbol{W}_K^T\boldsymbol{X}^T\boldsymbol{\varepsilon}) \\ &\quad + trace(\boldsymbol{\varepsilon}^T\boldsymbol{W}_K^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{W}_K\boldsymbol{\varepsilon}) \\ SS_E(K,J) &= trace(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T) - trace(\boldsymbol{X}\boldsymbol{W}_K\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T) - trace(\boldsymbol{W}_K^T\boldsymbol{X}^T\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T) \\ &\quad + trace(\boldsymbol{W}_K^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{W}_K\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T) \end{aligned} \tag{5.63}$$

Taking the expectation of equation (:

$$\begin{aligned} E[SS_E(K,J)] &= trace(\boldsymbol{\Lambda}_J) - trace(\boldsymbol{X}\boldsymbol{W}_K\boldsymbol{\Lambda}_J) - trace(\boldsymbol{W}_K^T\boldsymbol{X}^T\boldsymbol{\Lambda}_J) \\ &\quad + trace(\boldsymbol{W}_K^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{W}_K\boldsymbol{\Lambda}_J) \end{aligned} \tag{5.64}$$

Now, consider $E[SS_M(K,J)]$. Recall that $SS_M(K,J) = \left(\widehat{\boldsymbol{Y}}_K - \overline{\boldsymbol{Y}}\right)^T\left(\widehat{\boldsymbol{Y}}_K - \overline{\boldsymbol{Y}}\right)$.

Expanding the definition of $SS_M(K,J)$,

$$
\begin{aligned}
SS_M(K,J) &= \left(\widehat{\boldsymbol{Y}}_K - \overline{\boldsymbol{Y}}\right)^T \left(\widehat{\boldsymbol{Y}}_K - \overline{\boldsymbol{Y}}\right) \\
&= (\boldsymbol{X\beta} + \boldsymbol{XW}_K\,\boldsymbol{\varepsilon} - \boldsymbol{NX\beta} - \boldsymbol{N\varepsilon})^T (\boldsymbol{X\beta} + \boldsymbol{XW}_K\,\boldsymbol{\varepsilon} - \boldsymbol{NX\beta} - \boldsymbol{N\varepsilon}) \\
&= \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X\beta} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{XW}_K\,\boldsymbol{\varepsilon} - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{NX\beta} - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{N\varepsilon} \\
&+ \boldsymbol{\varepsilon}^T\boldsymbol{W}_K^T\boldsymbol{X}^T\boldsymbol{X\beta} + \boldsymbol{\varepsilon}^T\boldsymbol{W}_K^T\boldsymbol{X}^T\boldsymbol{XW}_K\,\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}^T\boldsymbol{W}_K^T\boldsymbol{X}^T\boldsymbol{NX\beta} - \boldsymbol{\varepsilon}^T\boldsymbol{W}_K^T\boldsymbol{X}^T\boldsymbol{N\varepsilon} \\
&- \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{NX\beta} - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{NXW}_K\,\boldsymbol{\varepsilon} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{NNX\beta} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{NN\varepsilon} \\
&- \boldsymbol{\varepsilon}^T\boldsymbol{NX\beta} - \boldsymbol{\varepsilon}^T\boldsymbol{NXW}_K\,\boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}^T\boldsymbol{NNX\beta} + \boldsymbol{\varepsilon}^T\boldsymbol{MN\varepsilon}
\end{aligned}
\tag{5.65}
$$

Note that $\boldsymbol{NN} = \boldsymbol{N}$. Also, note that

$$
\begin{aligned}
E[\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{XW}_K\,\boldsymbol{\varepsilon}] &= E[\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{N\varepsilon}] = E[\boldsymbol{\varepsilon}^T\boldsymbol{W}_K^T\boldsymbol{X}^T\boldsymbol{X\beta}] \\
&= E[\boldsymbol{\varepsilon}^T\boldsymbol{W}_K^T\boldsymbol{X}^T\boldsymbol{NX\beta}] = E[\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{NXW}_K\,\boldsymbol{\varepsilon}] \\
&= E[\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{NN\varepsilon}] = E[\boldsymbol{\varepsilon}^T\boldsymbol{NX\beta}] = E[\boldsymbol{\varepsilon}^T\boldsymbol{NNX\beta}] \\
&= 0
\end{aligned}
\tag{5.66}
$$

Taking the expectation of (5.65) and recalling the cyclic property of the trace yields:

$$
\begin{aligned}
E[SS_M(K,J)] &= \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X\beta} + 0 - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{NX\beta} - 0 + 0 + trace\left(\boldsymbol{W}_K^T\boldsymbol{X}^T\boldsymbol{XW}_K\,\boldsymbol{\Lambda}_J\right) \\
&\quad - 0 - trace(\boldsymbol{W}_K^T\boldsymbol{X}^T\boldsymbol{N\Lambda}_J) - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{NX\beta} - 0 + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{NX\beta} + 0 \\
&\quad - 0 - trace\left(\boldsymbol{NXW}_K\,\boldsymbol{\Lambda}_J\right) + 0 + trace\left(\boldsymbol{N\Lambda}_J\right) \\
&= \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X\beta} - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{NX\beta} + trace\left(\boldsymbol{W}_K^T\boldsymbol{X}^T\boldsymbol{XW}_K\,\boldsymbol{\Lambda}_J\right) \\
&\quad - trace(\boldsymbol{W}_K^T\boldsymbol{X}^T\boldsymbol{N\Lambda}_J) - trace\left(\boldsymbol{NXW}_K\,\boldsymbol{\Lambda}_J\right) + trace\left(\boldsymbol{N\Lambda}_J\right)
\end{aligned}
\tag{5.67}
$$

Now consider $2E[CP(K,J)]$. Recall that $2CP(K,J) = 2\left(\widehat{\boldsymbol{Y}}_K - \overline{\boldsymbol{Y}}\right)^T\left(\widetilde{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_K\right)$.

Expanding this definition yields:

$$\begin{aligned}
2CP(K,J) &= 2\left(\widehat{\boldsymbol{Y}}_K - \overline{\boldsymbol{Y}}\right)^T\left(\widetilde{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_K\right) \\
&= 2(\boldsymbol{X\beta} + \boldsymbol{XW}_K\,\boldsymbol{\varepsilon} - \boldsymbol{N\bar{X}\beta} - \boldsymbol{N\bar{\varepsilon}})^T(\boldsymbol{X\beta} + \boldsymbol{\varepsilon} - \boldsymbol{X\beta} - \boldsymbol{XW}_K\,\boldsymbol{\varepsilon}) \\
&= 2(\boldsymbol{X\beta} + \boldsymbol{XW}_K\,\boldsymbol{\varepsilon} - \boldsymbol{N\bar{X}\beta} - \boldsymbol{N\bar{\varepsilon}})^T(\boldsymbol{\varepsilon} - \boldsymbol{XW}_K\,\boldsymbol{\varepsilon}) \\
&= 2(\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{\varepsilon} - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{XW}_K\,\boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}^T\boldsymbol{W}_K^T\boldsymbol{X}^T\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}^T\boldsymbol{W}_K^T\boldsymbol{X}^T\boldsymbol{XW}_K\boldsymbol{\varepsilon} \\
&\quad - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{N\varepsilon} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{NXW}_K\,\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}^T\boldsymbol{N\varepsilon} + \boldsymbol{\varepsilon}^T\boldsymbol{NXW}_K\,\boldsymbol{\varepsilon})
\end{aligned} \tag{5.68}$$

Note that

$$E[\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{\varepsilon}] = E[\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{XW}_K\,\boldsymbol{\varepsilon}] = E[\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{N\varepsilon}] = E[\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{NXW}_K\,\boldsymbol{\varepsilon}] = 0 \tag{5.69}$$

Taking the expectation of equation ( and recalling the cyclic property of the trace:

$$\begin{aligned}
2E[CP(K,J)] &= 2\Big(0 - 0 + trace\left(\boldsymbol{W}_K^T\boldsymbol{X}^T\boldsymbol{\Lambda}_J\right) - trace\left(\boldsymbol{W}_K^T\boldsymbol{X}^T\boldsymbol{XW}_K\boldsymbol{\Lambda}_J\right) - 0 + 0 \\
&\quad - trace\left(\boldsymbol{N\Lambda}_J\right) + trace\left(\boldsymbol{NXW}_K\boldsymbol{\Lambda}_J\right)\Big) \\
&= 2trace\left(\boldsymbol{W}_K\boldsymbol{X}^T\boldsymbol{\Lambda}_J\right) - 2trace\left(\boldsymbol{W}_K\boldsymbol{X}^T\boldsymbol{XW}_K\boldsymbol{\Lambda}_J\right) \\
&\quad - 2trace\left(\boldsymbol{N\Lambda}_J\right) + 2trace\left(\boldsymbol{NXW}_K\boldsymbol{\Lambda}_J\right)
\end{aligned} \tag{5.70}$$

Substituting equations (, (5.67), and (5.70) into equation (5.62) and simplifying yields:

$$
\begin{aligned}
E[SS_T] = {} & trace(\boldsymbol{\Lambda}_J) - trace(\boldsymbol{X}\boldsymbol{W}_K\boldsymbol{\Lambda}_J) - trace(\boldsymbol{W}_K^T\boldsymbol{X}^T\boldsymbol{\Lambda}_J) \\
& + trace(\boldsymbol{W}_K^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{W}_K\boldsymbol{\Lambda}_J) + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{N}\boldsymbol{X}\boldsymbol{\beta} \\
& + trace(\boldsymbol{W}_K^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{W}_K\,\boldsymbol{\Lambda}_J) - trace(\boldsymbol{W}_K^T\boldsymbol{X}^T\boldsymbol{N}\boldsymbol{\Lambda}_J) \\
& - trace(\boldsymbol{N}\boldsymbol{X}\boldsymbol{W}_K\,\boldsymbol{\Lambda}_J) + trace(\boldsymbol{N}\boldsymbol{\Lambda}_J) + 2trace(\boldsymbol{W}_K^T\boldsymbol{X}^T\boldsymbol{\Lambda}_J) \\
& - 2trace(\boldsymbol{W}^T{}_K\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{W}_K\boldsymbol{\Lambda}_J) - 2trace(\boldsymbol{N}\boldsymbol{\Lambda}_J) \\
& + 2trace(\boldsymbol{N}\boldsymbol{X}\boldsymbol{W}_K\boldsymbol{\Lambda}_J) \\
& = (\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{I}-\boldsymbol{N})^T\boldsymbol{X}\boldsymbol{\beta} + trace(\boldsymbol{\Lambda}_J) - trace(\boldsymbol{N}\boldsymbol{\Lambda}_J) \\
& = (\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{I}-\boldsymbol{N})^T\boldsymbol{X}\boldsymbol{\beta} + n\hat{\sigma}_\delta^2(k) + \sum_{i=1}^{n} var(\tilde{y}_i) - trace(\boldsymbol{N}\boldsymbol{\Lambda}_J)
\end{aligned}
\tag{5.71}
$$

$$
\begin{aligned}
E[SS_T] = {} & (\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{I}-\boldsymbol{N})^T\boldsymbol{X}\boldsymbol{\beta} + SS(Model\ Error) + SS(Sampling\ Error) \\
& - trace(\boldsymbol{N}\boldsymbol{\Lambda}_J)
\end{aligned}
$$

The last line in equation (5.71) relates $E[SS_T]$ to the Gruber et al. [2007] pseudo ANOVA (Table 5.7). The second and third terms explain the variation due to model and sampling errors respectively, and are identical to those proposed by Gruber et al. [2007] (see Table 5.7). The first term describes the variation of the true model about its mean, and the fourth term is a correction for the errors in computing the sample mean. This correction term can be further decomposed. Let $\boldsymbol{T} = \boldsymbol{N}\boldsymbol{\Lambda}_J$. The $i^{th}$ diagonal element of $\boldsymbol{T}$, $t_{ii}$, is

$$t_{ii} = \frac{1}{n}\left(\hat{\sigma}_{\delta}^{2}(k) + var(\tilde{y}_i)\right) + \frac{1}{n}\sum_{j \neq i} \boldsymbol{\Lambda}_J(i,j) \tag{5.72}$$

where $\rho_{ij}$ is the cross-correlation of the sampling errors. The correction term in equation (5.71) is computed as

$$trace\left(\boldsymbol{N}\boldsymbol{\Lambda}_J\right) = \sum_{i=1}^{\mathrm{n}} t_{ii} = \hat{\sigma}_{\delta}^{2}(k) + \frac{1}{n}\sum_{i=1}^{n} var(\tilde{y}_i) + \frac{1}{n}\sum_{i=1}^{n}\sum_{j \neq i} \boldsymbol{\Lambda}_J(i,j) \tag{5.73}$$

Substituting equation (5.73) into equation (5.71) yields

$$\begin{aligned} E[SS_T] &= (\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{I}-\boldsymbol{N})\boldsymbol{X}\boldsymbol{\beta} + n\hat{\sigma}_{\delta}^{2}(k) + \sum_{i=1}^{n} var(\tilde{y}_i) - \hat{\sigma}_{\delta}^{2}(k) \\ &\quad - \frac{1}{n}\sum_{i=1}^{n} var(\tilde{y}_i) - \frac{1}{n}\sum_{i=1}^{n}\sum_{j \neq i} \boldsymbol{\Lambda}_J(i,j) \\ E[SS_T] &= (\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{I}-\boldsymbol{N})\boldsymbol{X}\boldsymbol{\beta} + (n-1)\hat{\sigma}_{\delta}^{2}(k) + \left(1-\frac{1}{n}\right)\sum_{i=1}^{n} var(\tilde{y}_i) \\ &\quad - \frac{1}{n}\sum_{i=1}^{n}\sum_{j \neq i} \boldsymbol{\Lambda}_J(i,j) \end{aligned} \tag{5.74}$$

Unfortunately, $\boldsymbol{\beta}$ is not know. If $\boldsymbol{b}_K$ is used instead, an estimate of $E[SS_T]$ is

$$\begin{aligned} \widehat{E[SS_T]} &= (\boldsymbol{X}\boldsymbol{b}_K)^T(\boldsymbol{I}-\boldsymbol{N})\boldsymbol{X}\boldsymbol{b}_K + (n-1)\hat{\sigma}_{\delta}^{2}(k) + \left(1-\frac{1}{n}\right)\sum_{i=1}^{n} var(\tilde{y}_i) \\ &\quad - \frac{1}{n}\sum_{i=1}^{n}\sum_{j \neq i} \boldsymbol{\Lambda}_J(i,j) \end{aligned} \tag{5.75}$$

Instead, suppose a *J* case analysis is used to both estimate the model parameters and to assess the precision of the model, the estimate of $E[SS_T]$ becomes

$$\widehat{E[SS_T]} = \left(\boldsymbol{X}\boldsymbol{b}_J\right)^T(\boldsymbol{I}-\boldsymbol{N})\boldsymbol{X}\boldsymbol{b}_J + (n-1)\hat{\sigma}_\delta^2(k) + \left(1-\frac{1}{n}\right)\sum_{i=1}^{n} var(\tilde{y}_i) - \frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}\boldsymbol{\Lambda}_J(i,j) \tag{5.76}$$

APPENDIX A

**Table A.1:** Censoring Decisions for analysis in Chapter 4, Part 1 of 7

| Site # | Site Name | POR | Type of Cens | 1-Day | 3-Day | 7-Day | 15-Day | 30-Day |
|---|---|---|---|---|---|---|---|---|
| 1 | Sacramento R Shasta Dam | 77 | EMA Cens/Zeros | 1 | 0 | 0 | 0 | 0 |
| | | | Additional Censored | 2 | 2 | 2 | 2 | 2 |
| | | | Total | 3 | 2 | 2 | 2 | 2 |
| 3 | Cottonwood Ck near Cottonwood | 68 | EMA Cens/Zeros | 1 | 1 | 1 | 1 | 1 |
| | | | Additional Censored | 0 | 0 | 0 | 0 | 0 |
| | | | Total | 1 | 1 | 1 | 1 | 1 |
| 4 | Cow Ck near Millville | 59 | EMA Cens/Zeros | 1 | 1 | 1 | 1 | 1 |
| | | | Additional Censored | 0 | 0 | 0 | 0 | 0 |
| | | | Total | 1 | 1 | 1 | 1 | 1 |
| 5 | Battle Ck below Coleman Hatch | 68 | EMA Cens/Zeros | 1 | 0 | 0 | 0 | 0 |
| | | | Additional Censored | 0 | 1 | 1 | 1 | 1 |
| | | | Total | 1 | 1 | 1 | 1 | 1 |
| 6 | Mill Ck near LosMolinos | 80 | EMA Cens/Zeros | 1 | 1 | 1 | 1 | 1 |
| | | | Additional Censored | 0 | 0 | 0 | 0 | 0 |
| | | | Total | 1 | 1 | 1 | 1 | 1 |
| 7 | Elder Ck near Paskenta | 60 | EMA Cens/Zeros | 0 | 0 | 1 | 1 | 1 |
| | | | Additional Censored | 1 | 1 | 0 | 0 | 0 |
| | | | Total | 1 | 1 | 1 | 1 | 1 |
| 8 | Thomes Ck at Paskenta | 76 | EMA Cens/Zeros | 1 | 1 | 1 | 1 | 1 |
| | | | Additional Censored | 0 | 0 | 0 | 0 | 0 |
| | | | Total | 1 | 1 | 1 | 1 | 1 |

**Table A.2:** Censoring Decisions for analysis in Chapter 4, Part 2 of 7

| Site # | Site Name | POR | Type of Cens | 1-Day | 3-Day | 7-Day | 15-Day | 30-Day |
|---|---|---|---|---|---|---|---|---|
| 9 | Deer Ck near Vina | 92 | EMA Cens/Zeros | 1 | 1 | 1 | 1 | 1 |
| | | | Additional Censored | 0 | 0 | 0 | 0 | 0 |
| | | | Total | 1 | 1 | 1 | 1 | 1 |
| 10 | BigChico Ck near Chico | 77 | EMA Cens/Zeros | 1 | 1 | 1 | 1 | 1 |
| | | | Additional Censored | 0 | 0 | 0 | 0 | 0 |
| | | | Total | 1 | 1 | 1 | 1 | 1 |
| 11 | Stony Ck at BlackButteDam | 66 | EMA Cens/Zeros | 1 | 1 | 1 | 1 | 1 |
| | | | Additional Censored | 0 | 0 | 0 | 0 | 0 |
| | | | Total | 1 | 1 | 1 | 1 | 1 |
| 12 | Butte Ck near Chico | 78 | EMA Cens/Zeros | 1 | 1 | 1 | 1 | 1 |
| | | | Additional Censored | 0 | 0 | 0 | 0 | 0 |
| | | | Total | 1 | 1 | 1 | 1 | 1 |
| 13 | Feather R At OrovilleDam | 107 | EMA Cens/Zeros | 0 | 0 | 0 | 0 | 0 |
| | | | Additional Censored | 0 | 0 | 0 | 0 | 0 |
| | | | Total | 0 | 0 | 0 | 0 | 0 |
| 14 | North Yuba at BullardsDam | 68 | EMA Cens/Zeros | 0 | 0 | 0 | 1 | 1 |
| | | | Additional Censored | 2 | 2 | 2 | 1 | 1 |
| | | | Total | 2 | 2 | 2 | 2 | 2 |
| 15 | Bear R near Wheatland | 103 | EMA Cens/Zeros | 0 | 0 | 1 | 1 | 1 |
| | | | Additional Censored | 1 | 1 | 0 | 0 | 0 |
| | | | Total | 1 | 1 | 1 | 1 | 1 |
| 16 | N Fork Cache Ck at IV Dam | 78 | EMA Cens/Zeros | 1 | 1 | 1 | 1 | 1 |
| | | | Additional Censored | 3 | 3 | 3 | 3 | 3 |
| | | | Total | 4 | 4 | 4 | 4 | 4 |

**Table A.3:** Censoring Decisions for analysis in Chapter 4, Part 3 of 7

| Site # | Site Name | POR | Type of Cens | 1-Day | 3-Day | 7-Day | 15-Day | 30-Day |
|---|---|---|---|---|---|---|---|---|
| 17 | American R at FairOaks | 104 | EMA Cens/Zeros | 0 | 0 | 0 | 1 | 1 |
| | | | Additional Censored | 1 | 1 | 1 | 0 | 0 |
| | | | Total | 1 | 1 | 1 | 1 | 1 |
| 18 | Kings R at Pine Flat Dam | 113 | EMA Cens/Zeros | 0 | 0 | 0 | 0 | 0 |
| | | | Additional Censored | 0 | 0 | 0 | 0 | 0 |
| | | | Total | 0 | 0 | 0 | 0 | 0 |
| 19 | SanJoaquin R at Friant Dam | 105 | EMA Cens/Zeros | 0 | 0 | 0 | 0 | 0 |
| | | | Additional Censored | 0 | 0 | 0 | 0 | 0 |
| | | | Total | 0 | 0 | 0 | 0 | 0 |
| 20 | Chowchilla R at Buch Dam | 80 | EMA Cens/Zeros | 1 | 0 | 0 | 0 | 0 |
| | | | Additional Censored | 0 | 1 | 1 | 1 | 1 |
| | | | Total | 1 | 1 | 1 | 1 | 1 |
| 23 | DelPuerto Ck near Patterson | 44 | EMA Cens/Zeros | 1 | 0 | 0 | 0 | 1 |
| | | | Additional Censored | 0 | 1 | 1 | 1 | 0 |
| | | | Total | 1 | 1 | 1 | 1 | 1 |
| 24 | Merced R at Exchequer Dam | 107 | EMA Cens/Zeros | 1 | 1 | 1 | 1 | 1 |
| | | | Additional Censored | 0 | 0 | 0 | 0 | 0 |
| | | | Total | 1 | 1 | 1 | 1 | 1 |
| 25 | Tuolumne R at DonPedroDam | 112 | EMA Cens/Zeros | 0 | 0 | 0 | 0 | 0 |
| | | | Additional Censored | 0 | 0 | 0 | 0 | 0 |
| | | | Total | 0 | 0 | 0 | 0 | 0 |
| 26 | Stanislaus R at MelonesDam | 93 | EMA Cens/Zeros | 0 | 0 | 0 | 0 | 0 |
| | | | Additional Censored | 1 | 1 | 1 | 1 | 1 |
| | | | Total | 1 | 1 | 1 | 1 | 1 |

**Table A.4:** Censoring Decisions for analysis in Chapter 4, Part 4 of 7

| Site # | Site Name | POR | Type of Cens | 1-Day | 3-Day | 7-Day | 15-Day | 30-Day |
|---|---|---|---|---|---|---|---|---|
| 28 | Duck Ck near Farmington | 30 | EMA Cens/Zeros | 1 | 1 | 0 | 0 | 0 |
| | | | Additional Censored | 0 | 0 | 1 | 1 | 1 |
| | | | Total | 1 | 1 | 1 | 1 | 1 |
| 30 | Calaveras R at Hogan Dam | 96 | EMA Cens/Zeros | 1 | 1 | 1 | 1 | 1 |
| | | | Additional Censored | 0 | 0 | 0 | 0 | 0 |
| | | | Total | 1 | 1 | 1 | 1 | 1 |
| 31 | Mokelumne R at Camanche Dam | 104 | EMA Cens/Zeros | 0 | 0 | 0 | 0 | 0 |
| | | | Additional Censored | 0 | 0 | 0 | 0 | 0 |
| | | | Total | 0 | 0 | 0 | 0 | 0 |
| 32 | Cosumnes R at Michigan Bar | 101 | EMA Cens/Zeros | 1 | 1 | 1 | 1 | 1 |
| | | | Additional Censored | 0 | 0 | 0 | 0 | 0 |
| | | | Total | 1 | 1 | 1 | 1 | 1 |
| 33 | Fresno R near Knowles | 76 | EMA Cens/Zeros | 0 | 0 | 0 | 0 | 0 |
| | | | Additional Censored | 0 | 0 | 0 | 0 | 0 |
| | | | Total | 0 | 0 | 0 | 0 | 0 |
| 34 | S Yuba R atJones Bar | 57 | EMA Cens/Zeros | 0 | 0 | 0 | 1 | 1 |
| | | | Additional Censored | 1 | 1 | 1 | 0 | 0 |
| | | | Total | 1 | 1 | 1 | 1 | 1 |
| 35 | M Yuba R below OurHouseDam | 37 | EMA Cens/Zeros | 0 | 0 | 0 | 0 | 0 |
| | | | Additional Censored | 0 | 0 | 0 | 0 | 0 |
| | | | Total | 0 | 0 | 0 | 0 | 0 |
| 36 | Kaweah R at Terminus Dam | 50 | EMA Cens/Zeros | 0 | 0 | 0 | 0 | 0 |
| | | | Additional Censored | 0 | 0 | 0 | 0 | 0 |
| | | | Total | 0 | 0 | 0 | 0 | 0 |

**Table A.5:** Censoring Decisions for analysis in Chapter 4, Part 5 of 7

| Site # | Site Name | POR | Type of Cens | 1-Day | 3-Day | 7-Day | 15-Day | 30-Day |
|---|---|---|---|---|---|---|---|---|
| 37 | Tule R at Success Dam | 50 | EMA Cens/Zeros | 0 | 0 | 0 | 0 | 0 |
| | | | Additional Censored | 0 | 0 | 0 | 0 | 0 |
| | | | Total | 0 | 0 | 0 | 0 | 0 |
| 38 | Kern R Isabella Dam | 116 | EMA Cens/Zeros | 0 | 0 | 0 | 0 | 0 |
| | | | Additional Censored | 0 | 0 | 0 | 0 | 0 |
| | | | Total | 0 | 0 | 0 | 0 | 0 |
| 39 | Mill Ck near Piedra | 52 | EMA Cens/Zeros | 0 | 0 | 0 | 0 | 0 |
| | | | Additional Censored | 0 | 0 | 0 | 0 | 0 |
| | | | Total | 0 | 0 | 0 | 0 | 0 |
| 40 | Dry Ck near Lemoncove | 50 | EMA Cens/Zeros | 0 | 0 | 1 | 1 | 0 |
| | | | Additional Censored | 1 | 1 | 0 | 0 | 1 |
| | | | Total | 1 | 1 | 1 | 1 | 1 |
| 41 | Deer Ck near Fount Spr | 41 | EMA Cens/Zeros | 0 | 0 | 0 | 0 | 0 |
| | | | Additional Censored | 0 | 0 | 0 | 0 | 0 |
| | | | Total | 0 | 0 | 0 | 0 | 0 |
| 42 | White R near Ducor | 46 | EMA Cens/Zeros | 0 | 0 | 0 | 0 | 0 |
| | | | Additional Censored | 0 | 0 | 0 | 0 | 0 |
| | | | Total | 0 | 0 | 0 | 0 | 0 |
| 43 | Cache Ck at Clear Lake | 87 | EMA Cens/Zeros | 1 | 1 | 1 | 1 | 1 |
| | | | Additional Censored | 3 | 3 | 3 | 3 | 3 |
| | | | Total | 4 | 4 | 4 | 4 | 4 |
| 44 | Putah Ck at Mont Dam | 78 | EMA Cens/Zeros | 1 | 1 | 2 | 2 | 2 |
| | | | Additional Censored | 11 | 11 | 9 | 9 | 9 |
| | | | Total | 12 | 12 | 11 | 11 | 11 |

**Table A.6:** Censoring Decisions for analysis in Chapter 4, Part 6 of 7

| Site # | Site Name | POR | Type of Cens | 1-Day | 3-Day | 7-Day | 15-Day | 30-Day |
|---|---|---|---|---|---|---|---|---|
| 45 | M Fork Eel R near DosRios | 43 | EMA Cens/Zeros | 1 | 1 | 1 | 1 | 1 |
| | | | Additional Censored | 0 | 0 | 0 | 0 | 0 |
| | | | Total | 1 | 1 | 1 | 1 | 1 |
| 46 | S Fork Eel R near Miranda | 68 | EMA Cens/Zeros | 1 | 1 | 1 | 1 | 1 |
| | | | Additional Censored | 0 | 0 | 0 | 0 | 0 |
| | | | Total | 1 | 1 | 1 | 1 | 1 |
| 47 | Mad R above Ruth Res | 28 | EMA Cens/Zeros | 0 | 0 | 0 | 0 | 0 |
| | | | Additional Censored | 0 | 0 | 0 | 0 | 0 |
| | | | Total | 0 | 0 | 0 | 0 | 0 |
| 48 | E Fork Rus R near Calpella | 67 | EMA Cens/Zeros | 1 | 1 | 1 | 1 | 1 |
| | | | Additional Censored | 0 | 0 | 0 | 0 | 0 |
| | | | Total | 1 | 1 | 1 | 1 | 1 |
| 49 | Salinas R near Pozo | 41 | EMA Cens/Zeros | 0 | 0 | 0 | 0 | 0 |
| | | | Additional Censored | 0 | 0 | 0 | 0 | 0 |
| | | | Total | 0 | 0 | 0 | 0 | 0 |
| 50 | Arroyo Seco near Soledad | 107 | EMA Cens/Zeros | 0 | 1 | 1 | 1 | 1 |
| | | | Additional Censored | 1 | 0 | 0 | 0 | 0 |
| | | | Total | 1 | 1 | 1 | 1 | 1 |
| 51 | Salmon R at SomesBar | 84 | EMA Cens/Zeros | 1 | 1 | 1 | 1 | 1 |
| | | | Additional Censored | 0 | 0 | 0 | 0 | 0 |
| | | | Total | 1 | 1 | 1 | 1 | 1 |
| 52 | SantaCruz Ck near SantaYnez | 67 | EMA Cens/Zeros | 2 | 2 | 1 | 0 | 0 |
| | | | Additional Censored | 4 | 3 | 4 | 5 | 4 |
| | | | Total | 6 | 5 | 5 | 5 | 4 |

**Table A.7:** Censoring Decisions for analysis in Chapter 4, Part 7 of 7

| Site # | Site Name | POR | Type of Cens | 1-Day | 3-Day | 7-Day | 15-Day | 30-Day |
|---|---|---|---|---|---|---|---|---|
| 53 | Salsipuedes Ck near Lompoc | 67 | EMA Cens/Zeros | 0 | 0 | 0 | 0 | 0 |
| | | | Additional Censored | 0 | 0 | 0 | 0 | 0 |
| | | | Total | 0 | 0 | 0 | 0 | 0 |
| 54 | Trinity R above CoffeeCk | 51 | EMA Cens/Zeros | 0 | 1 | 1 | 1 | 1 |
| | | | Additional Censored | 1 | 0 | 0 | 0 | 0 |
| | | | Total | 1 | 1 | 1 | 1 | 1 |
| 55 | Scott R near FortJones | 67 | EMA Cens/Zeros | 1 | 1 | 1 | 1 | 1 |
| | | | Additional Censored | 0 | 0 | 0 | 0 | 0 |
| | | | Total | 1 | 1 | 1 | 1 | 1 |